BRNO UNIVERSITY OF TECHNOLOGY
VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

FACULTY OF INFORMATION TECHNOLOGY
FAKULTA INFORMAČNÍCH TECHNOLOGIÍ

DEPARTMENT OF COMPUTER SYSTEMS
ÚSTAV POČÍTAČOVÝCH SYSTÉMŮ

BIOINFORMATICS TOOLS FOR SEQUENCE AND STRUCTURAL DATA ANALYSIS

BIOINFORMATICKÉ NÁSTROJE PRO ANALÝZU SEKVENČNÍCH A STRUKTURNÍCH DAT

HABILITATION THESIS
HABILITAČNÍ PRÁCE

AUTHOR **Ing. TOMÁŠ MARTÍNEK, Ph.D.**
AUTOR PRÁCE

BRNO 2022

Abstract

Understanding the nature of living organisms is one of the central tasks of biology. In recent decades, there have been significant advances in this field, thanks in particular to new technologies that allow us to obtain vast amounts of information at the molecular level. However, the quantity and speed of the new data acquisition are so high that it is no longer possible to analyze it manually or to use simple statistical methods. As a result, there is increasing pressure to develop effective bioinformatics tools that help process this information in an automated and accurate way. This habilitation thesis summarizes newly developed algorithms and tools in three areas of bioinformatics in which the author participated. First, the triplex and pqsfinder tools for searching specific secondary structures in DNA, such as triplexes and quadruplexes, are presented. Both of these tools stand out for their ability to detect even non-perfect sequences involving different types of defects, which are also observed in real experiments. Then, newly developed tools in the field of protein engineering are presented to help design new proteins with desired properties. Specifically, these tools include SoluProt for protein solubility prediction, EnzymeMiner for mining enzymes of interest from large databases and prioritizing them, HotSpot Wizard for identifying protein positions suitable for mutagenesis, and FireProt for automated design of thermostable proteins. Finally, new tools and approaches to analyze repetitive regions of the genome are presented. In particular, a tool for detecting distant or novel Insertion Sequence elements in assembled prokaryotic genomes and a new approach for analyzing and visualizing satellite DNA directly from sequencing data.

Keywords

DNA secondary structures, repetitive DNA sequences, protein engineering, bioinformatics

Abstrakt

Pochopení podstaty živých organismů je jednou z centrálních úloh biologie. V posledních desetiletích byl v této oblasti zaznamenám významný posun a to zejména díky novým technologiím, které nám umožňují získávat obrovské množství informací na molekulární úrovni. Biologové již nejsou schopni takové množství dat analyzovat ručně nebo s využitím jednoduchých statistických metod. Při své práci se proto neobejdou bez pomoci efektivních bioinformatických nástrojů. Tato habilitační práce sumarizuje nově vyvinuté algoritmy a nástroje ve třech oblastech bioinformatiky, na kterých se autor podílel. Nejprve budou prezentovány nástroje triplex a pqsfinder pro vyhledávání specifických sekundárních struktur v DNA jako jsou triplexy a kvadruplexy. Oba tyto nástroje vynikají svou schopností detekovat i neperfektní sekvence zahrnující různé typy defektů, které jsou pozorovány i v reálných experimentech. Následně budou prezentovány nově vyvinuté nástroje v oblasti proteinového inženýrství, které pomáhají při designu nových proteinů s požadovanými vlastnostmi. Konkrétně se jedná o nástroje: SoluProt pro predikci solubility proteinu, EnzymeMiner pro dolování zájmových enzymů z rozsáhlých databází a jejich prioritizaci, Hotspot Wizard pro identifikaci pozic proteinu vhodných k mutagenezi a FireProt pro automatizovaný návrh teplotně stabilních proteinů. Na závěr budou prezentovány nové nástroje a přístupy v oblasti analýzy repetitivních oblastí genomu. Konkrétně bude představen nástroj pro detekci vzdálených nebo nových Insertion Sequence elementů v sestavených genomech prokaryot a nový přístup pro analýzu a vizualizaci satelitní DNA přímo ze sekvenačních dat.

Klíčová slova

Sekundární struktury DNA, opakující se sekvence DNA, proteinové inženýrství, bioinformatika

Reference

MARTÍNEK, Tomáš. *Bioinformatics tools for sequence and structural data analysis.* Brno, 2022. Habilitation thesis. Brno University of Technology, Faculty of Information Technology.

Bioinformatics tools for sequence and structural data analysis

Declaration

Hereby I declare that this habilitation's thesis was prepared as an original author's work. All the relevant information sources, which were used during preparation of this thesis, are properly cited and included in the list of references.

. . . .
Tomáš Martínek
November 17, 2022

Acknowledgements

I want to thank Dr. Matej Lexa, who brought me into the field of bioinformatics. His original ideas and friendly approach have always been a great inspiration for me during my research. Furthermore, I would like to thank the entire team of Loschmidt laboratories and especially Prof. Jiří Damborský for his expert guidance in protein engineering. I will never forget his helpful and dedicated approach and incredible diligence. Furthermore, I would like to thank my colleagues and collaborators from the Institute of Biophysics of the Czech Academy of Sciences. In particular, Dr. Marie Brázdová and Assoc. Prof. Eduard Kejnovský for the valuable advice they provided me during my research in the field of DNA sequence analysis. I would also like to thank my Ph.D. students: Jiří Hon, Miloš Musil, and Janka Puterová, for their cooperation in the research and the enormous effort they have devoted to developing these tools. Without their help, these outputs would not have been produced. Last but not least, I would like to thank my family for the support and patience they gave me not only during the writing of this thesis but especially throughout the research. In conclusion, I would like to thank the team of people who take care of the Metacentre's computing resources funded under the project "e-Infrastruktura CZ" (e-INFRA CZ LM2018140) supported by the Ministry of Education, Youth and Sports of the Czech Republic. Without these computing resources' support, many of this thesis's outputs would be very difficult to produce.

Contents

1 Introduction **5**

2 Secondary DNA structures **7**

2.1 Introduction . . . 7

2.2 State of the art . . . 8

2.2.1 Research objectives . . . 9

2.3 Research summary . . . 9

2.3.1 Identification of potential triplex-forming sequences . . . 9

2.3.2 Identification of potential quadruplex-forming sequences . . . 10

2.3.3 Additional biological studies . . . 12

2.3.4 List of publications . . . 14

2.4 Conclusions . . . 19

2.4.1 Future work . . . 19

2.5 Research impact . . . 19

3 Protein engineering **21**

3.1 Introduction . . . 21

3.2 Rational selection methods . . . 24

3.2.1 State of the art . . . 24

3.2.2 Research summary . . . 26

3.3 Rational protein design . . . 28

3.3.1 State of the art . . . 28

3.3.2 Research summary . . . 30

3.4 List of publications . . . 33

3.5 Conclusions . . . 37

3.5.1 Future work . . . 37

3.6 Research impact . . . 38

4 Repetitive DNA sequences **41**

4.1 Introduction . . . 41

4.1.1 Tandem repeats . . . 41

4.1.2 Transposable elements . . . 41

4.1.3 Study of repetitive sequences . . . 42

4.2 State of the art . . . 44

4.2.1 Assembly-based approaches . . . 44

4.2.2 Assembly-free approaches . . . 44

4.2.3 Research objectives . . . 45

4.3 Research summary . . . 46

4.3.1 digIS - novel approach for detection of distant Insertion Sequence elements . . . 46
4.3.2 Novel approach for detailed analysis of satellite DNA . . . 46
4.3.3 Additional studies . . . 47
4.3.4 List of publications . . . 49
4.4 Conclusions . . . 52
4.4.1 Future work . . . 52
4.5 Research impact . . . 52

Bibliography **55**

A Included Papers **79**
A.1 Paper I . . . 79
A.2 Paper II . . . 89
A.3 Paper III . . . 93
A.4 Paper IV . . . 101
A.5 Paper V . . . 113
A.6 Paper VI . . . 139
A.7 Paper VII . . . 147
A.8 Paper VIII . . . 155
A.9 Paper IX . . . 163
A.10 Paper X . . . 171
A.11 Paper XI . . . 193
A.12 Paper XII . . . 211

Chapter 1

Introduction

Bioinformatics is an interdisciplinary field of science that helps biologists solve complex problems using computers. These are primarily tasks in molecular biology that are characterized by large amounts of data and the complexity of its processing and visualization. Bioinformatics is, therefore, very closely related to biology, chemistry, physics, and computer science. It helps, for example, in the fields of analysis and processing of genomic data (Genomics), protein sequences (Proteomics), the study of complex biological processes (System biology), or the creation of simulation models (Molecular dynamics). Bioinformatics is such a broad field that it now encompasses almost anything related to biology and computer science.

The author of this thesis became familiar with bioinformatics through the design of digital circuits for hardware acceleration of selected algorithms for the analysis of biological sequences, which was carried out in collaboration with Dr. Matej Lexa from the Masaryk University. In the follow-up research, the author moved further towards developing software tools that fall into three specific subareas of bioinformatics. This thesis aims to present the results achieved in this follow-up research.

The first part of the thesis (Chapter 2) deals with the area of DNA secondary structures, such as hairpins, triplexes, or quadruplexes. They represent an alternative to the canonical (double helix) DNA, and their formation is conditioned by the specific nucleotide sequence and the physicochemical environment in which the molecule is located. The study of DNA secondary structures is of great interest to biologists, as it appears that these structures may have a significant impact on several biological processes, including the regulation of gene expression, mutagenesis, and the development of various diseases. To understand these complex processes, biologists need effective bioinformatics tools to detect and search for these secondary structures in DNA.

Therefore, in the first part of this work, we focused on designing new algorithms for detecting triplex and quadruplex forming sequences in DNA. Compared to existing tools, the developed outputs stand out for their ability to search even structures containing a certain degree of defects. The developed tools were subsequently used in several follow-up biological studies, where we analyzed the occurrence of triplex-forming sequences in the human genome or the interactions of these structures with one of the key tumor suppressor proteins – p53. This work was done in collaboration with the Institute of Biophysics of the Czech Academy of Sciences and the team around Dr. Marie Brázdová.

The second part of the thesis (Chapter 3) is devoted to protein engineering, which deals with modifying existing proteins to improve their valuable properties such as activity, stability, or selectivity. Such modified proteins are then attractive targets for pharmaceutical and industrial applications. Successful applications include drug design, biofuel production, detergents, waste treatment, food processing, the paper industry, and many others. Therefore, this part of the thesis focuses on developing new or improving existing approaches in this area. Specifically, we have been designing algorithms and tools for mining enzymes of interest from large biological databases, predicting protein solubility, predicting protein positions suitable for mutagenesis, and predicting multi-point mutations to increase protein stability. The tools developed are among the best in their category, as evidenced by their massive use by the scientific community and industry. This work was done in collaboration with Loschmidt laboratories and the team of Prof. Jiří Damborský.

The last third part of the thesis (Chapter 4) combines, to some extent, the author's previous experience in the field of genomics and proteomics. It deals with the analysis of genomic regions containing a large number of repetitive sequences. It has been shown that these regions are part of the genomes of most living organisms. For example, they make up more than two-thirds of the human genome, and their representation in plants tend to be much higher. These regions of the genome were originally referred to as junk DNA. However, an increasing number of studies have demonstrated their importance in various biological processes, including chromosome organization and rearrangements, the control of telomere elongation, or modulation of gene expression. To further understand the role of these repetitive regions, it is essential to provide the scientific community with efficient tools to search and analyze them in both assembled genomes and sequencing data.

Therefore, we have developed a new tool to search for Insertion Sequence elements (ISE) in assembled prokaryotic genomes. Compared to existing approaches, this tool stands out for detecting even putative novel ISE families. The field of tools for analyzing and quantifying repetitive sequences directly from sequencing data has been further expanded with a new way of processing satellite DNA sequences and their visualization. The developed technique was applied to the analysis of the seabuckthorn (Hippophae rhamnoides) genome and contributed to an interesting discovery regarding the size of X and Y chromosomes. In addition, a study analyzing the relationships between transposable elements and quadruplexes was conducted. The results of the study revealed several interesting findings regarding the occurrence of quadruplexes within or in the vicinity of TEs, including their successful experimental evaluation in vitro. This work was again done in collaboration with the Institute of Biophysics of the Czech Academy of Sciences and the team of Assoc. Prof. Eduard Kejnovský.

The thesis is presented as a set of articles. Each of the three areas is presented separately and includes an introduction to the field, state of the art, a definition of the objectives, and a summary of the results achieved. Copies of the published articles on which the thesis is based are then included in Appendix A.

Chapter 2

Secondary DNA structures

2.1 Introduction

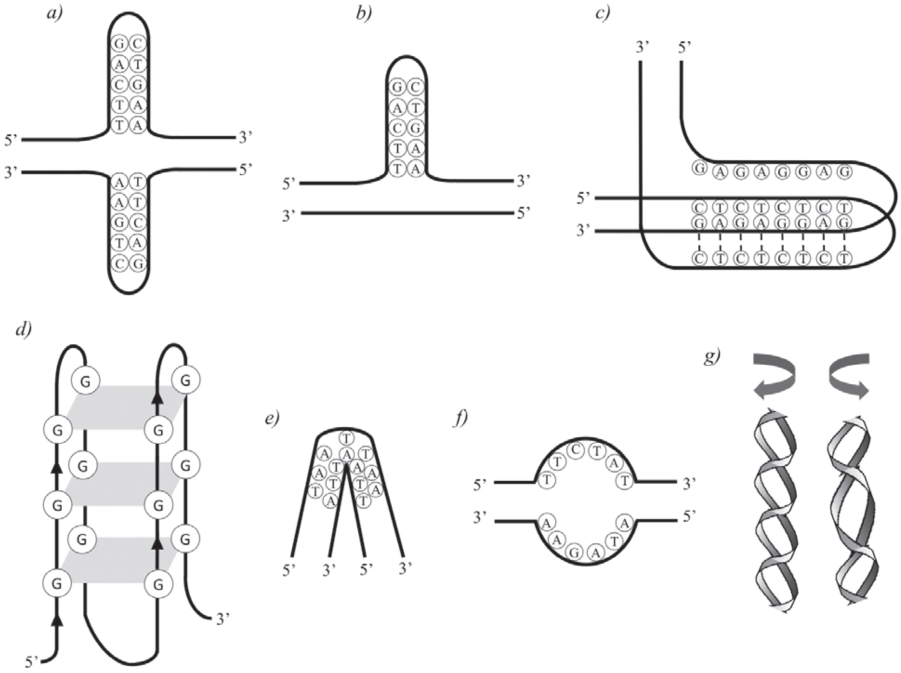
DNA is commonly known as a molecule made up of a pair of complementary strands composed of basic building blocks called nucleotides (Adenine, Guanine, Thymine, and Cytosine). Due to the presence of the complementary strand and the linkages between Adenine and Thymine or Guanine and Cytosine (known as Watson-Crick base pairing), DNA acquires its typical double helix structure. However, DNA does not always have to occupy only a double helix structure. At certain locations and under certain conditions, specific structures can occur on DNA in the form of hairpins, triplexes, quadruplexes, and other non-canonical DNA conformations (see Figure 2.1).

The formation of these structures is conditioned by a specific nucleotide sequence at a given DNA location, the bases of which form additional bonds within one or more strands. This behavior is best seen in cruciform structures (see Figure 2.1a), where complementary Watson-Crick bonds are formed within both the forward and reverse DNA strands to stabilize the structure. In the case of triplexes, the bonds between the three strands are stabilized by combining Hoogsteen and Watson-Crick base pairing (see Figure 2.1c). The quadruplex structure is formed even between a quartet of strands composed of Guanine sequences (see Figure 2.1d). The more layers of Guanines are present, the more stable the resulting structure is.

A specific nucleotide sequence is a necessary but insufficient condition for forming a non-canonical DNA structure. The physico-chemical environment in which the molecule is located or the additional stresses exerted on individual strands (e.g., DNA supercoiling) is also crucial. Thus, depending on the environmental conditions, the DNA molecule may adopt one of several stable conformations.

The study of these secondary structures is of great interest to biologists, as it appears that these structures can have a significant impact on many biological processes. For example, most of the observed hairpins and triplexes suggest roles in mutagenesis, recombination, and gene regulation. Non-B DNA structures have been shown to cause deletions, expansions and translocations in both prokaryotes and eukaryotes [158]. Their distribution is not random and often colocalizes with sites of chromosomal breakage [233]. Triplex structures can block the replication fork and result in double-stranded breaks [66]. In some cases, the mutagenesis induced by such sequences is enhanced by their transcription [21], possibly via transcriptional arrest. Also, quadruplexes (G4) are involved in mutagenesis and disease [13]. They are implicated in several genome-wide processes, mostly as positive or

Figure 2.1: Examples of non-B (non-canonical) DNA structures. (a) Cruciform; (b) hairpin; (c) triplex or H-DNA (shown is the Y*R:Y structure formed by Hoogsteen hydrogen bonding); (d) intramolecular tetraplex, G4-DNA or G-quadruplex with diagonal loop; (e) bent DNA; (f) unwound DNA; (g) comparison between B-DNA (left, clockwise) and Z- or left-handed DNA (right, anticlockwise). The figure was adapted from [143].



negative transcription regulators [165]. They may be dispersed into critical locations of the genome by the activity of transposable elements [120].

To understand these complex processes, it is therefore very important for biologists to have effective bioinformatics tools to search for these secondary structures in DNA.

2.2 State of the art

Several studies deal with the design of algorithms for detecting potential non-canonical DNA structures. However, many of these algorithms are very simplistic and often only search for structures without defects, for example, perfect hairpins, triplexes, or quadruplexes. However, scientific studies are increasingly showing that real secondary structures contain defects in the form of mutations. Despite these defects, they can be stable and provide important biological functions. For example, numerous papers have reported the existence of imperfect triplexes [132, 169, 223]. In recent years, different in vitro experiments have also confirmed the existence of imperfect G4s [138]. Imperfect G4s have also been explored in silico by molecular dynamics [210].

Simplistic tools usually fail to detect these structures, limiting the ability of biologists to study them. In the case of triplexes, existing tools are often based on homopurine and homopyrimidine tracts, which are most appropriate for detecting perfect triplexes [78]. Another work [40] created a web-based catalog of non-B DNA sequences in major mammalian genomes. Their definition of triplex covers the most stable canonical triplexes made of G.GC/A.AT and C.GC/T.AT triplets but leaves little room for possible errors. More complex sequence-structure relationships of triplexes were brought into a small number of computational tools for identifying relevant sequences in genomes. For example, Schroth and Ho [179] analyzed the occurrence of inverted and mirror repeats, and Hoyne et al. [93] studied the Escherichia coli genome (E.coli) for intrastrand triplex sequences.

Similarly, the most commonly used algorithms to search for quadruplexes are based on a simple folding rule representing four runs of guanines separated by relatively short loops (or spacers). These include quadparser [94], QGRS Mapper [57, 107], and Quadfinder [177]. The folding rule used is usually of the form G{3,6}.{1,8}G{3,6}.{1,8}G{3,6}.{1,8}G{3,6} reflecting the fact that potential quadruplex-forming sequences (PQS) with short loops and four perfect G runs form the most stable G4s in vitro. These tools consider only sequences that match the sequence formula perfectly.

New tools for the prediction of imperfect G4s have begun to be developed. Such tools include TetraplexFinder/QuadBase2 [63], ImGQfinder [209], and G4Hunter [19]. For example, TetraplexFinder considers potential bulges of defined length in runs of three guanines. In contrast, ImGQfinder considers the possibility of a single bulge or mismatch in a wider variety of guanine run lengths. Finally, G4Hunter does not define individual defect types but uses a simple encoding and statistics over a sliding window that can accommodate different types of defects.

2.2.1 Research objectives

This work aims to design and implement new and efficient bioinformatics tools for detecting non-canonical DNA secondary structures that would also consider different types of defects, allowing biologists to obtain more accurate information for their studies.

2.3 Research summary

2.3.1 Identification of potential triplex-forming sequences

We have developed a new algorithm for the detection of triplex-forming sequences in DNA that also considers different types of defects between Hoogsteen and Watson-Crick base pairing compared to existing tools. The proposed algorithm is based on the dynamic programming technique, widely used in bioinformatics, for example, for pairwise sequence alignment [190]. Thanks to this technique, the developed tool can detect mutations in the searched sequences in the form of a character substitution, insertion, or deletion. An advantage is also that the scoring function of the algorithm can be adjusted to best meet the specific characteristics of the triplex structure.

For example, based on the studies of Rathinavelan and Yathindra [160]; Thenmalarchelvi and Yathindra [203], which discuss different combinations of disorders in the backbone of triplets, we decided to divide the individual triplets into isomorphic groups. As triplets from one group are more likely to form stable triplexes than other sequences, a special penalty to the scoring function for changing the isomorphic group was added. In addition, we verified the ability of triplets to form bonds and their membership in isomorphic groups using molecular simulation and the AmberTools [145]. The resulting scoring function was further adjusted based on an analysis of published experimentally verified structures of perfect and imperfect triplexes.

The developed algorithm was tested on real genomes (E.coli and human) and on randomized sequences into which sequences of real triplexes from the non-B DNA database were inserted [40]. The results show that the developed algorithm achieves high processing speed and, simultaneously, shows increased sensitivity in finding triplexes with different types of defects.

More detailed information about the developed algorithm and the achieved results can be found in the original version of the manuscript in Appendix A.1, published in the Bioinformatics journal.

Bioconductor package

Bioinformatics tools often suffer because the original authors stop working on the topic once the publication is released. The softwares created age gradually to the point where they become non-functional. One way to solve this problem is to release the source code of the tools as open and allow others to develop and maintain them. The usability and availability of the tools are then increased if they become part of a frequently used platform where they can interoperate with other tools and form part of, for example, a more complex genomic pipeline.

One such open platform is Bioconductor[1], which currently contains more than two thousand bioinformatic software packages. This platform is based on the R language but also allows the user to integrate code written in other programming languages, such as C/C++. The platform also includes basic object class definitions for different areas of biological data analysis. For example, classes like *GRanges* are often used for genomic data to identify specific regions in genomes. Using these predefined classes, it is easy to link different applications within the platform to more complex pipelines.

For the above reasons, we decided to transfer our implemented algorithm for searching triplex-forming sequences to Bioconductor. We changed the inputs and outputs of the tool to objects of type *DNAString* and *GRanges*. All necessary modifications were made according to this platform's requirements and passed comprehensive acceptance tests. In addition to the triplex search, we also implemented 3D visualization of the found outputs and their export to standardized annotation formats such as GFF3.

More detailed information about the package and its capabilities can be found in the original version of the manuscript in Appendix A.2, published in the Bioinformatics journal.

2.3.2 Identification of potential quadruplex-forming sequences

In this area, we have developed a new tool, pqsfinder, for detecting potential quadruplex-forming sequences (PQS) in DNA, which also considers non-perfect quadruplexes (G4) in the search. Based on a study of the literature and experimentally verified G4s, two basic types of defects were identified: mismatches and bulges (insertions inside G-runs).

We then designed an algorithm that first identifies four consecutive imperfect G-run sequences (G run quartet). Subsequently, it examines the potential of such a G-run quartet to form a stable G4 and reports a corresponding quantitative score. Internally, the tool relies on a suitable combination of regular expression and backtracking to find all overlapping quadruplexes. The backtracking procedure increases the computational complexity of the search but allows us to model the competition between overlapping PQS rigorously. The user can list all or only the non-overlapping PQS with the highest score.

[1] `https://www.bioconductor.org/`

To score the quadruplexes found, we adopted an approach where the score is modular and obtained by adding scores representing the binding affinities of smaller regions within the G4. This approach has already been proven to work for simpler DNA structures, such as nucleic acid duplexes and hairpins [175, 235]. The first part of the scoring scheme quantifies the quality of individual G runs. It awards the PQS with a score for each G-tetrad stacking and penalizes mismatches and bulges in G runs. Based on the available literature, bulges and long loops are considered to be strong destabilizers of G4s and do not expect more than a few of these imperfections to be possible simultaneously.

The resulting scoring function is based on several penalization constants that are difficult to determine analytically. Therefore, we decided to use a technique to train these penalties based on experimental data. Consequently, two datasets were constructed based on the results of the literature study. The first dataset consisted of experimentally validated G4 sequences obtained from published studies, 392 sequences in total (Lit392). Unfortunately, this set displayed several shortcomings, such as being unbalanced in terms of positive and negative samples. Furthermore, it contained a small number of samples representing only a fraction of possible G4 conformations, including a limited number of errors in the form of mismatches, bulges, and different loop lengths.

To create the second dataset, we took advantage of the unique outputs of the work of Chambers et al. [43], where the authors introduced a new technique for high-throughput sequencing of G4 structures called G4-Seq. The technique detects noisy sequences that emerge when treating DNA samples with Kþ or PDS (pyridostatin, a chemical G4 stabilizer). In other words, the authors experimentally measured individual pieces of human DNA and their ability to form G4. As a result of this technology, the authors released a track (in BED format) that shows the propensity of reference Human DNA sequence (hg19) to form G4s.

This unique and extensive dataset includes a tremendous amount of information about potential G4 sequences, including various deformations and their effect on stability. Therefore, we decided to use this second dataset primarily to train the penalty constants of the pqsfinder tool. The first dataset (Lit392) containing a limited number of G4s compiled from published studies was used as a test set.

For pqsfinder training and parameter-space exploration, we took advantage of the genetic algorithm implemented in the R package GA [181]. To evaluate fitness, we calculated Pearson's correlation coefficient between the maximum score vector generated by pqsfinder and the vector representing the propensity of the sequence to form G4 from the G4-seq training set. The basic idea behind this fitness function is: the higher the correlation coefficient between pqsfinder score and G4 propensity level, the better the prediction of putative G4 structures will be.

Comparison results showed that on the Lit392 dataset, pqsfinder significantly outperformed existing tools in Matthews' correlation coefficient (a suitable metric when the test set is unbalanced). We also prepared a second dataset (from G4-Seq) for testing while keeping the sequences for training and testing of the tool strictly separated. Again, the pqsfinder tool performed significantly better than competing tools QGRS Mapper and G4Hunter.

Similar to the triplex search algorithm, we implemented pqsfinder as a package for the R Bioconductor environment. In addition to searching for PQS, the software offers its visualization and export to standardized annotation formats such as GFF3.

More detailed information about the developed algorithm, package, and results of comparison against competing tools can be found in the original version of the manuscript in Appendix A.3, published in the Bioinformatics journal.

2.3.3 Additional biological studies

Distribution of triplex-forming sequences in human genome

Using the developed triplex search tool, we performed a lookup and detailed analysis of the occurrence of potential triplex-forming sequences (PTS) within the human genome. Only PTS occurrences with a P-value <0.05 were considered for the study. Then the regions where the PTSs occurred were analyzed, and their numbers were compared against those we would observe if they were hypothetically randomly distributed in the genome.

The analysis showed that PTS was found in higher numbers in gene promoters, introns, and intergenic regions. In contrast, a reduced abundance of PTS was observed in coding sequences, 5'UTR and 3'UTR regions. This finding is consistent with basic assumptions about the occurrence of these secondary structures. That is, they are more likely to occur in the regulatory areas of the genome and not to interfere with coding segments.

Given the higher prevalence of PTS in intergenic areas, a detailed analysis of these regions was needed. Intergenic areas are primarily occupied by repetitive sequences of different types, ranging from short tandem repeats to transposon sequences of several thousand base pairs. We analyzed the positions of the PTS concerning the different classes and families of these repetitive elements. This experiment showed an increased occurrence of PTS in SVA, Alu[2], and low complexity regions.

Upon first inspection, it becomes clear that most of the associations mentioned above are caused by the presence of the polyA[3] tail in SINE elements. Because the poly-A tail is mainly described as a feature circumventing the problematic polyadenylation in RNA polymerase III transcripts [172], there is a possibility that these sequences do not form any functionally or evolutionarily meaningful DNA structures, such as triplexes. On closer inspection, however, we have noticed that the same classes of repeats are also enriched for other PTS sequences, raising the possibility that triplex formation plays a biological role in the repeat life cycles at the DNA level. This could also mean a dual role for the Alu poly-A tail. For example, Dewannieux and Heidmann [62] mention a 15-50 nucleotide range for the increasing effect of the poly-A tail, a range that also coincides with cited oligonucleotide lengths for successful triplex formation [37].

More detailed information about the study, the results obtained, and their discussion can be found in the original version of the manuscript in Appendix A.4, published at the International Conference on Bioinformatics Models, Methods and Algorithms.

Interactions between p53 and triplex-forming sequences

Previous studies of PTS occurrences within the human genome have shown that these sequences are most commonly found in promoter regions. This observation is consistent with studies demonstrating that triplex-forming sequences play important roles within various regulatory processes [214]. However, it is necessary to focus on a specific biological process, the corresponding set of genes or regulatory proteins such as transcription factors, to demonstrate such a role for the triplex. It is also necessary to demonstrate that such behavior occurs in vitro (in a glass) or even in vivo (in a living organism). In a follow-up study, we, therefore, collaborated with colleagues from the Institute of Biophysics of the

[2] Alu sequences are short non-autonomous retrotransposons (SINE) driven by the L1 LINE element protein machinery [61]. SVA elements are evolutionarily related to SINE and Alu sequences; therefore, the increased occurrence of PTS in these elements is not surprising.

[3] polyA sequence is a good candidate for triplex formation

Czech Academy of Sciences to investigate the role of triplexes and their effect on the activity of p53, one of the most studied proteins affecting key biological processes within the cell.

The p53 protein is a nuclear protein (393 amino acids) that is the product of one of the key tumor suppressor genes, TP53. The protein functions as a transcription factor and has a role of a DNA damage sensor in the cell. In the physiological state, the p53 protein is inactive. When DNA damage occurs, a signaling cascade is induced, resulting in activation of the p53 protein. This further causes cell cycle arrest in the G1 phase, giving the cell time to repair. If DNA repair is successful, the cell can resume the cell cycle. Otherwise, the cell induces apoptosis (cell death) [142].

The p53 protein is known to bind to genes in two ways: (i) it recognizes a DNA-specific consensus sequence in the form of 5'-PuPuPuC(A/T)(T/A)GPyPyPy-3' (CON) separated by 0±13 bp [71], and (ii) it can also bind to specific secondary structures such as cruciforms [96], DNA loops [198], or G4s [2].

In this study, we were the first to analyze the interactions between the p53 protein and DNA containing triplex-forming sequences in vitro and in cells. Using luciferase reporter assay in two different cell systems, we demonstrated that T.A.T triplex-forming sequences in front of CON, enhanced promoter activation by p53. Interestingly, the reporter vector containing only the T.A.T triplex-forming sequence was repressed by p53 protein. Both these effects suggested that T.A.T triplex-forming sequences have the potential to influence transcription in both directions. We assume that the positioning of T.A.T triplex on the promoter region facilitates p53 recognition and transcription of genes.

Based on these findings, we then performed an in-silico analysis of the human genome to find all genes that contain CON sequences in their promoter together with the T.A.T. triplex and thus may be significantly affected by p53 protein and related biological processes. As an output, we received 43 promoters of candidate p53 target genes with at least one CON and a T.A.T triplex with a poly(A/T) run longer than 40 bp. Using STRING-db and enrichment analysis, we identified 16 genes/proteins out of 43 that are strongly linked in terms of function and most closely match the GO term „regulation of signal transduction".

More detailed information about the study conducted, the results obtained, and their discussion can be found in the original version of the manuscript in Appendix A.5, published in the Plos One journal.

2.3.4 List of publications

Publication I

Title	A dynamic programming algorithm for identification of triplex-forming sequences
Authors	LEXA Matej, MARTÍNEK Tomáš, BURGETOVÁ Ivana, KOPEČEK Daniel, and BRÁZDOVÁ Marie
Abstract	**Motivation:** Current methods for identification of potential triplex-forming sequences in genomes and similar sequence sets rely primarily on detecting homopurine and homopyrimidine tracts. Procedures capable of detecting sequences supporting imperfect, but structurally feasible intramolecular triplex structures are needed for better sequence analysis. **Results:** We modified an algorithm for detection of approximate palindromes, so as to account for the special nature of triplex DNA structures. From available literature, we conclude that approximate triplexes tolerate two classes of errors. One, analogical to mismatches in duplex DNA, involves nucleotides in triplets that do not readily form Hoogsteen bonds. The other class involves geometrically incompatible neighboring triplets hindering proper alignment of strands for optimal hydrogen bonding and stacking. We tested the statistical properties of the algorithm, as well as its correctness when confronted with known triplex sequences. The proposed algorithm satisfactorily detects sequences with intramolecular triplex-forming potential. Its complexity is directly comparable to palindrome searching. **Availability:** Our implementation of the algorithm is available at `http://www.fi.muni.cz/~lexa/triplex` as source code and a web-based search tool. The source code compiles into a library providing searching capability to other programs, as well as into a stand-alone command-line application based on this library.
Journal	Bioinformatics, vol. 27, num. 18, 2011 Journal impact factor: 5.468, Q1
Citations	13 (WoS without self-citations)
Author's contribution	Algorithm design and implementation, manuscript writing (partially).
Manuscript	Appendix A.1

Publication II

Title	Triplex: an R/Bioconductor package for identification and visualization of potential intramolecular triplex patterns in DNA sequences
Authors	HON Jiří, MARTÍNEK Tomáš, RAJDL Kamil, and LEXA Matej
Abstract	**Motivation:** Upgrade and integration of triplex software into the R/Bioconductor framework. **Results:** We combined a previously published implementation of a triplex DNA search algorithm with visualization to create a versatile R/Bioconductor package 'triplex'. The new package provides functions that can be used to search Bioconductor genomes and other DNA sequence data for occurrence of nucleotide patterns capable of forming intramolecular triplexes (H-DNA). Functions producing 2D and 3D diagrams of the identified triplexes allow instant visualization of the search results. Leveraging the power of Biostrings and GRanges classes, the results get fully integrated into the existing Bioconductor framework, allowing their passage to other Genome visualization and annotation packages, such as GenomeGraphs, rtracklayer or Gviz. **Availability:** R package 'triplex' is available from Bioconductor (`bioconductor.org`).
Journal	Bioinformatics, vol. 29, num. 15, 2013 Journal impact factor: 4.621, Q1
Citations	13 (WoS without self-citations)
Author's contribution	Optimization of the original algorithm, consultation on the design and implementation of the tool, testing of the final package, manuscript writing (partially).
Manuscript	Appendix A.2

Publication III

Title pqsfinder: an exhaustive and imperfection-tolerant search tool for potential quadruplex-forming sequences in R

Authors HON Jiří, MARTÍNEK Tomáš, ZENDULKA Jaroslav, and LEXA Matej

Abstract

Motivation: G-quadruplexes (G4s) are one of the non-B DNA structures easily observed in vitro and assumed to form in vivo. The latest experiments with G4-specific antibodies and G4-unwinding helicase mutants confirm this conjecture. These four-stranded structures have also been shown to influence a range of molecular processes in cells. As G4s are intensively studied, it is often desirable to screen DNA sequences and pinpoint the precise locations where they might form.

Results: We describe and have tested a newly developed Bioconductor package for identifying potential quadruplex-forming sequences (PQS). The package is easy-to-use, flexible and customizable. It allows for sequence searches that accommodate possible divergences from the optimal G4 base composition. A novel aspect of our research was the creation and training (parametrization) of an advanced scoring model which resulted in increased precision compared to similar tools. We demonstrate that the algorithm behind the searches has a 96% accuracy on 392 currently known and experimentally observed G4 structures. We also carried out searches against the recent G4-seq data to verify how well we can identify the structures detected by that technology. The correlation with pqsfinder predictions was 0.622, higher than the correlation 0.491 obtained with the second best G4Hunter.

Availability: `http://bioconductor.org/packages/pqsfinder/` This paper is based on pqsfinder-1.4.1.

Journal Bioinformatics, vol. 33, num. 21, 2017
Journal impact factor: 5.481, Q1

Citations 52 (WoS without self-citations)

Author's contribution Studying state-of-the-art, consulting on algorithm design and implementation, training penalization constants, preparing datasets, testing the final package, writing manuscript (partially).

Manuscript Appendix A.3

Publication IV

Title	Uneven distribution of potential triplex sequences in the human genome: In silico study using the R/Bioconductor package triplex
Authors	LEXA Matej, MARTÍNEK Tomáš, and BRÁZDOVÁ Marie
Abstract	Eukaryotic genomes are rich in sequences capable of forming non-B DNA structures. These structures are expected to play important roles in natural regulatory processes at levels above those of individual genes, such as whole genome dynamics or chromatin organization, as well as in processes leading to the loss of these functions, such as cancer development. Recently, a number of authors have mapped the occurrence of potential quadruplex sequences in the human genome and found them to be associated with promoters. In this paper, we set out to map the distribution and characteristics of potential triplex-forming sequences (PTS) in the human genome sequence. Using the R/Bioconductor package triplex, we found these sequences to be excluded from exons, while present mostly in a small number of repetitive sequence classes, especially short sequence tandem repeats (microsatellites), Alu and combined elements, such as SVA. We also introduce a novel way of classifying potential triplex sequences, using a lexicographically minimal rotation of the most frequent k-mer to assign class membership automatically. Members of such classes typically have different propensities to form parallel and antiparallel intramolecular triplexes (H-DNA). We observed an interesting pattern, where the predicted third strands of antiparallel H-DNA were much less likely to contain a deletion than their duplex structural counterpart than were their parallel versions.
Conference	International Conference on Bioinformatics Models, Methods and Algorithms, 2014
Citations	1 (WoS without self-citations)
Author's contribution	Implementation of individual experiments and their evaluation, manuscript writing (partially).
Manuscript	Appendix A.4

Publication V

Title	p53 Specifically Binds Triplex DNA In Vitro and in Cells
Authors	BRÁZDOVÁ Marie, TICHÝ Vlastimil, HELMA Robert, BAŽANTOVÁ Pavla, POLÁŠKOVÁ Alena, KREJČÍ Aneta, PETR Marek, NAVRÁTILOVÁ Lucie, TICHÁ Olga, NEJEDLÝ Karel, BENNINK Martin L., SUBRAMANIAM Vinod, BÁBKOVÁ Zuzana, MARTÍNEK Tomáš, LEXA Matej, and ADÁMIK Matej
Abstract	Triplex DNA is implicated in a wide range of biological activities, including regulation of gene expression and genomic instability leading to cancer. The tumor suppressor p53 is a central regulator of cell fate in response to different type of insults. Sequence and structure specific modes of DNA recognition are core attributes of the p53 protein. The focus of this work is the structure-specific binding of p53 to DNA containing triplex-forming sequences in vitro and in cells and the effect on p53-driven transcription. This is the first DNA binding study of full-length p53 and its deletion variants to both intermolecular and intramolecular T.A.T triplexes. We demonstrate that the interaction of p53 with intermolecular T.A.T triplex is comparable to the recognition of CTG-hairpin non-B DNA structure. Using deletion mutants we determined the C-terminal DNA binding domain of p53 to be crucial for triplex recognition. Furthermore, strong p53 recognition of intramolecular T.A.T triplexes (H-DNA), stabilized by negative superhelicity in plasmid DNA, was detected by competition and immunoprecipitation experiments, and visualized by AFM. Moreover, chromatin immunoprecipitation revealed p53 binding T.A.T forming sequence in vivo. Enhanced reporter transactivation by p53 on insertion of triplex forming sequence into plasmid with p53 consensus sequence was observed by luciferase reporter assays. In-silico scan of human regulatory regions for the simultaneous presence of both consensus sequence and T.A.T motifs identified a set of candidate p53 target genes and p53-dependent activation of several of them (ABCG5, ENOX1, INSR, MCC, NFAT5) was confirmed by RT-qPCR. Our results show that T.A.T triplex comprises a new class of p53 binding sites targeted by p53 in a DNA structure-dependent mode in vitro and in cells. The contribution of p53 DNA structure-dependent binding to the regulation of transcription is discussed.
Journal	PLoS ONE, vol. 11, num. 12, 2016 Journal impact factor: 2.766, Q1
Citations	11 (WoS without self-citations)
Author's contribution	In-silico analysis of promoter regions, search for triplexes and P53 consensus sequences.
Manuscript	Appendix A.5

2.4 Conclusions

In the field of DNA secondary structures, we have designed and implemented new tools for the detection of triplex and quadruplex-forming sequences in DNA. Compared to existing tools, both of them are superior in finding not only patterns for typical structures but even those with a certain level of defects. This functionality allows biologists to explore the area of non-canonical DNA structures better, gain a broader view of their occurrence in real sequences, and better estimate their influence in various biological processes.

For wider availability of the developed tools, we also implemented them for the R Bioconductor environment. This approach allows their integration with other tools for DNA sequence analysis, including automation of experimental data processing in the form of genomic pipelines.

We have also used the developed tools in several follow-up biological studies. In particular, we analyzed the occurrence of triplex-forming sequences in the human genome, which confirmed their dominant occurrence in promoter regions. We have also investigated the interactions of these structures with the p53 protein, demonstrating these structures' significant role in the regulation of gene transcription.

2.4.1 Future work

Although the developed tools for searching triplex and quadruplex-forming sequences take into account different types of defects and push the quality of outputs one step further, we are still far from considering them (or any other tool) as final. While we have tried to incorporate as much information as possible into the tools, many factors still have not been considered. For example, in the case of triplexes we do not consider: the competition between alternative structures [168], fourth strand (the strand which is not part of the predicted triplex), effects of C+ distribution [95, 183] and other distortions caused by electrostatic forces [99, 200]. Most of these factors depend non-trivially on the environment [150]. Since the algorithm does not consider the environment, it is limited to sequence-coded effects only.

Similarly, the relationship between DNA sequence and G4 structure is very complex for quadruplexes. Despite our ability to model this relationship directly at the molecular level, using, for example, molecular dynamics and AmberTools [174], this approach is computationally demanding, and the accuracy of the state-of-the-art force fields is still limited. Therefore, existing tools for G4 prediction (including our pqsfinder) use much simpler models and currently represent a reasonable compromise between accuracy and computational complexity. In the future, we can expect a gradual development and improvement of these tools.

2.5 Research impact

Both developed tools for searching triplex and quadruplex-forming sequences are widely used by the research community, as evidenced by the statistics on the number of downloads of these tools from the Bioconductor environment[4]. These show that the triplex tool has been downloaded approximately 8.2 thousand times (from unique IP addresses) since its launch in 2013. In the case of the pqsfinder tool, it has accumulated approximately 4.6

[4] `http://bioconductor.org/packages/stats/bioc/triplex/`, `http://bioconductor.org/packages/stats/bioc/pqsfinder/`

thousand downloads since 2016. However, these numbers should be taken with caution as they may include downloads based on package updates. In addition, the statistics for unique IP addresses are calculated separately for each year. Next, the unique IP address type metric may not reflect the actual number of users, as a single user may use multiple IP addresses over time. A more accurate estimate might be on the order of hundreds of unique users. Which would also correspond to the number of citations received. In the case of the triplex tool, this is 26 citations, and for the pqsfinder tool it is 52. (It is assumed that not every user completes his/her work to publication.)

In the case of additional biological studies, the impact was smaller. The publication analyzing the distribution of triplexes in the human genome has a low citation rate, probably due to the choice of a conference instead of a journal. On the other hand, a study on the interaction of p53 protein and triplex-forming sequences has received 12 citations.

Chapter 3

Protein engineering

3.1 Introduction

Proteins are molecules that play a key role in all living organisms. For example, they form the building blocks of muscles and tissues, serve to transmit signals, transport molecules, or catalyze biochemical reactions. The detailed study of the structure, function, and interactions between proteins and other molecules is the subject of a field called structural biology.

Protein engineering is a discipline that deals with the subsequent modification of known wild-type proteins to obtain a new, improved protein function [33]. Elementary modifications include inserting, removing, or substituting specific amino acids in protein chains. However, sufficient knowledge of the structure and function of a given protein is necessary to perform these modifications effectively. Protein engineering is, therefore, very closely linked to structural biology and other disciplines.

The central interest of protein engineering is in proteins that catalyze chemical reactions, called enzymes or biocatalysts. By appropriate modification of these enzymes, it is possible to obtain, for example, a variant that can accelerate or inhibit a given chemical reaction or a variant that can function in various environments, e.g., with a higher temperature or a different pH. Such modified enzymes are then attractive targets for pharmaceutical and industrial applications. Successful examples of their use include drug design, biofuel production, detergents, waste treatment, food processing, paper industry, and many others [45].

In the field of protein engineering, two basic strategies are used to create new enzymes: (i) directed evolution and (ii) rational design. The basic idea of directed evolution is to mimic the evolutionary process in nature. A large number of mutations are randomly generated in the gene of the protein of interest. These modified genes are then inserted into expression systems, where they are used to create the corresponding proteins. Finally, a screening process is carried out to verify the individual variants of the protein in terms of the desired properties (e.g., activity, stability, selectivity), and the final product is selected.

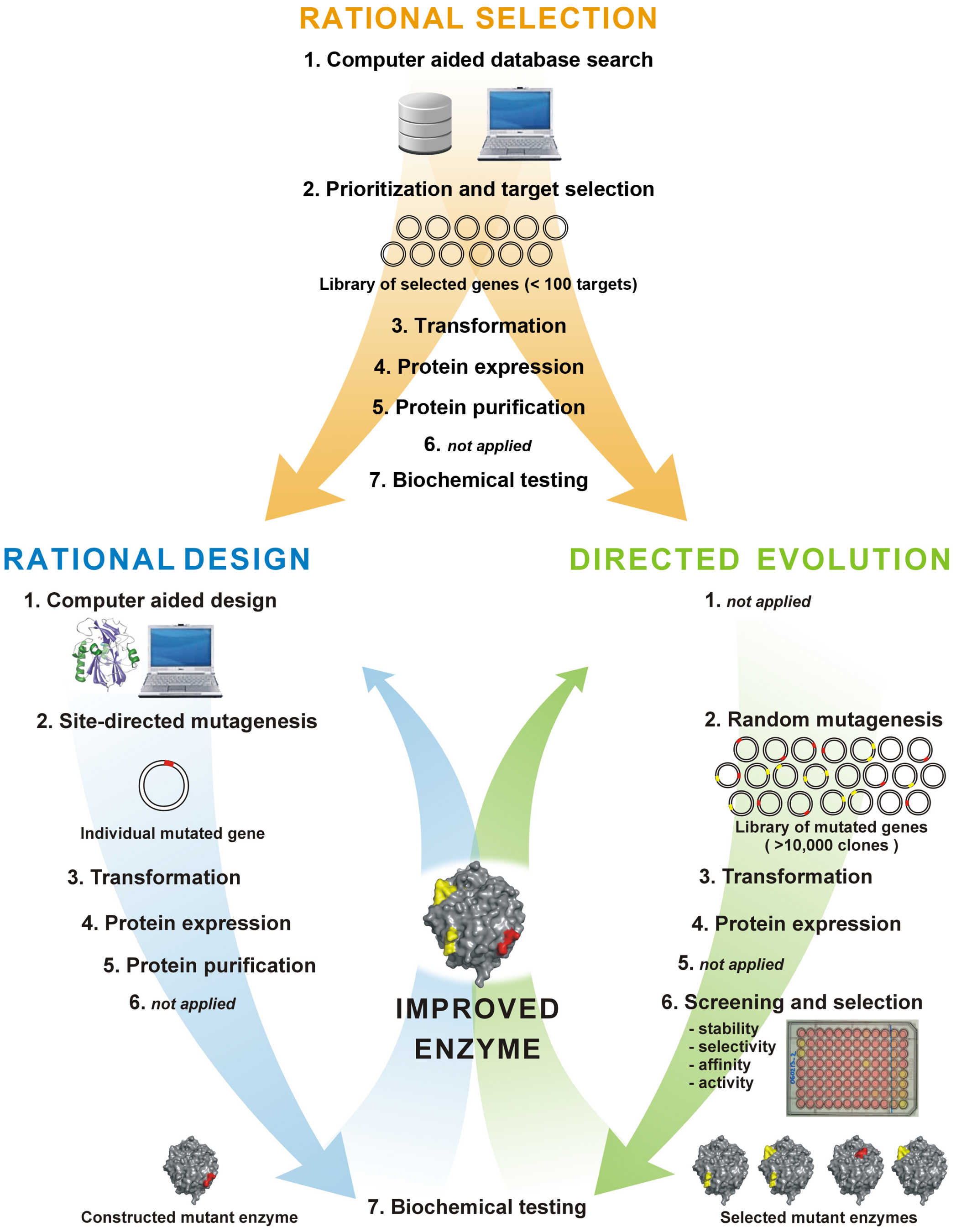
In contrast, rational design is based on making targeted changes instead of random ones. Protein engineers verify only specific mutations based on a deep knowledge of protein structure and function and using computational tools. The advantage of this technique is that it is significantly less demanding in terms of experimental work and less expensive. Instead of a large library of mutants, only a few targeted mutations are produced, which are referred to as smart libraries. On the other hand, this method requires deep knowledge of the protein of interest, including its 3D structure, obtained, e.g., through X-ray crystallography [191]. Mutations are then designed based on extensive computational analyses accompanied by molecular-level simulations [102, 6].

These two basic protein engineering strategies have been complemented in recent years by a third strategy called rational selection. This strategy is based on the hypothesis that similar proteins from different organisms perform a similar function but may have interesting properties because they have evolved independently in different organisms and thus have had to adapt to different environments. An illustrative example is thermophilic bacteria living in high-temperature environments. The protein of interest found in these bacteria is likely to be more thermally stable. This protein variant can be directly produced or used as a starting model for further mutations. Therefore, the cornerstone of rational selection is the combination of expert knowledge with database searches based on sequence similarity, complemented by computational analysis.

All these three strategies can be effectively combined, as illustrated in Figure 3.1. Rational selection can be a source of suitable genes for both rational design and directed evolution techniques. In addition, rational design can be used to accurately identify protein subregions suitable for directed evolution.

In the field of protein engineering, we have developed new tools and methods for rational selection and rational design strategies. For better clarity, the outputs produced are divided into two separate sections.

Figure 3.1: Protein engineering methods. The goal of protein engineering is to design a protein with improved properties, usually an enzyme for the catalysis of biochemical reactions. The rational design uses previous expert knowledge and computational simulations to design individual improved protein variants. Directed evolution relies on random mutagenesis and high-throughput screening of generated gene libraries. Rational selection provides alternative starting proteins based on computer-aided database mining for both rational design and directed evolution. The figure was adapted from the previous work by Damborský [56].



3.2 Rational selection methods

3.2.1 State of the art

Mining enzymes of interest

Thanks to significant advances in sequencing technologies, we are experiencing a tremendous increase in the amount of biological data. It is no longer possible to manually process this amount of data, extract all genes and proteins, and experimentally verify their function. Therefore, the vast majority of data is annotated and classified automatically using complex genomic pipelines such as the GenBank Annotation Pipeline [124].

The current knowledge about existing proteins is accumulated in large protein databases such as UniProtKB [1] and NCBI Protein [176]. These databases are often divided into two parts: (i) experimentally obtained and manually annotated proteins and (ii) computationally predicted proteins. According to the UniProtKB database—a comprehensive, high-quality, and freely accessible resource of protein sequence and functional information, approximately 500,000 experimentally verified proteins represent less than 0.3% of all deposited proteins. The remaining 209 million computationally predicted proteins represent a huge source of potentially interesting and diverse proteins for both basic science and industrial applications.

We can search these large databases based on metadata (extracted from the annotation pipeline) or sequence similarity. Since the automated annotation and metadata assignments are based on a tiny number of experimentally validated and characterized proteins (0.3%), their accuracy is limited for now [164]. Therefore, many studies and bioinformatics tools use the second option, i.e., data mining based on sequence similarity. In this case, the main input is the sequence of the protein of interest and the target of the search is the set of proteins with the highest similarity. Many different algorithms and tools have been developed for this purpose. Some of the most well-known ones include: BLAST [7], HMMER [67], UBLAST [69], RAPSearch2 [234], MMseqs2 [197], and DIAMOND [36].

One of the key problems of current methods for sequence-based protein search is the huge number of results found. Depending on the given parameters and the family size of the protein of interest, thousands to tens of thousands of hits can be obtained. Since only units or small tens of candidate sequences can usually be selected for experimental evaluation, these hits must be extensively filtered, categorized, and prioritized based on various criteria. Unfortunately, to date, we are not aware of a tool that can automate these steps for the field of protein engineering and search for enzymes of interest. Thus, many departments have to deal with laborious manual work or focus on developing their own software.

Prediction of protein solubility

In the subsequent steps of the rational selection strategy, it is crucial to produce the selected proteins easily. For these purposes, so-called expression systems and a technique called recombinant protein expression (RPE) [53] are used. The basic principle of this technique is to prepare a gene capable of encoding the protein of interest and insert it into the genome of a living organism (usually into a specific plasmid of E.coli or other simple organisms). Suitable conditions are then induced to trigger the expression of the gene, and the living organism starts producing corresponding proteins using standard biological processes such as transcription and translation. Finally, the produced proteins are extracted from the expressing organism and isolated for subsequent screening.

Trying to create a protein using a foreign organism involves many risks. Firstly, the protein being created usually comes from another organism, which generally means a different environment (temperature, pH range, etc.). The expression system may lack key components, such as chaperones, to properly fold the generated protein into its native 3D structure. Finally, the protective system of the organism used for expression may recognize the produced protein as foreign and target it for degradation. These may be just some of the few reasons why the protein of interest fails to be produced.

In general, the ability of a protein to fold into a proper crystal structure in a given solution is referred to as protein solubility [10]. This key property is dependent not only on extrinsic factors (e.g., the environment of the expression system) but also on the particular amino acid sequence. For the time being, the exact relationship between an amino acid sequence and its solubility is unknown. Still, tools for predicting solubility based on the sequence are already being developed, as well as various strategies for increasing solubility using appropriate mutations.

The first methods for solubility prediction include simple multi-parameter models or regression analysis. This category includes, for example, the Wilkinson-Harrison model and its extended variants [219, 58, 65], as well as the Protein-Sol tool [88] or the calculation of a solubility-weighted index [28].

Machine learning-based methods that predict the global solubility level of a protein based on a set of extracted sequence and physicochemical features are considered more advanced. Tools in this category include SOLPro [129], PROSSO II [189], and ESPRESSO [89].

More advanced types of tools can calculate not only the global solubility level but also which parts of the sequence are more sensitive in terms of solubility, calculating the so-called solubility profile. This property is typical for the tools such as ccSOL [5] and CamSol [193].

Finally, methods based on convolutional neural networks and deep learning techniques have also been developed. This category includes DeepSol [105] and SKADE [159].

Unfortunately, these tools' prediction accuracy is insufficient and often overestimated. An independent study by Chang et al. reported a large drop of 10-20% in the accuracy of existing tools when evaluated using a larger test set [44]. While most tools report accuracies around 70-80%, measurements on an independent dataset reveal prediction accuracy close to the 50% threshold (equivalent to random prediction). The reasons for the observed low accuracy may be different. The most commonly cited include insufficient training and test set preparation, or an insufficient number of samples available for training compared to all known proteins. Last but not least, we still do not know the exact relationship between the amino acid sequence and its solubility. It is estimated that there is, for example, a close link to the mechanisms of protein folding, representing a very complex process. Thus, the sets of basic sequence and physicochemical features used so far may be insufficient to generalize such a complex process.

Research objectives

This work is focused on improving methods in the area of rational selection. The main goal is to develop a new tool for mining enzymes of interest from large databases, which will include efficient filtering of the hits found and their prioritization in terms of different criteria, especially solubility, which is crucial for protein production.

This task, therefore, includes a part focusing on improving existing methods for protein solubility prediction. The aim is to address the shortcomings of current tools, focus on the precise preparation of datasets, and explore the possibility of more advanced predicted features.

3.2.2 Research summary

Protein solubility

We have created a new solubility prediction tool called SoluProt. This tool is based on machine learning, specifically a gradient boosting machine learning technique [77]. The input is the amino acid sequence, and the output is the predicted solubility level of the protein.

When creating the tool, we focused on the precise preparation of the training and testing dataset. We prepared the training set based on information extracted from the TargetTrack database [27]. This database maintains the outputs of several Protein Structure Initiative (PSI) projects that aimed to collect a large amount of information about protein 3D structures. Before obtaining information about the ternary structure, it is necessary to produce the protein, usually using RPE. Therefore, as a by-product of these PSI experiments, information about the protein's solubility is also available.

Although the TargetTrack database is structured, the solubility information is not available directly. We had to deduce it by analyzing the recorded states that the protein undergoes during its production and crystallization. The output is then usually just binary information - soluble/insoluble protein. Moreover, key information about the expression system used in each study had to be parsed and inferred from the text records in a complex way.

We were the first to create a carefully curated dataset for solubility predictors based on experimentally obtained data from the TargetTrack database. The created dataset is balanced not only in terms of positive and negative samples but also with respect to the representation of different protein lengths. In addition, it only includes samples related to the E. coli-based expression system, thus removing bias from the data. In total, the dataset generated consists of 11,436 samples.

Based on communication with researchers from the North East Structural Consortium (NESG), we also gained access to unique experimental data from protein solubility measurements. In contrast to the information from the TargetTrack database, solubility was measured at up to five levels, with each measurement repeated several times to verify its validity. Furthermore, all measurements were always performed using the same procedure. It is, therefore, one of the most accurate sources of protein solubility data available today. In total, this dataset contains 3,100 samples.

In addition to preparing the datasets, we also worked intensively on the selection of suitable features for machine learning. As a result, we selected 96 features falling into different groups, including complex predicted features such as: (i) average flexibility as computed by DynaMine [52], (ii) secondary structure content as predicted by FELLS [148], (iii) average disorder as predicted by ESPRITZ [213], and (iv) content of amino acids in transmembrane helices as predicted by TMHMM [111].

To compare the newly developed SoluProt tool with competing tools, we used the TargetTrack dataset for training and the NESG dataset for testing. This approach creates a fairer comparison since most existing ML-based tools have been trained on the TargetTrack data. The more accurate NESG data was then used as a test set to verify the quality of each tool.

Similar to the independent study by Chang et al. [44], we found that the accuracy of the tools on an independent dataset ranges from 50-60%. Our SoluProt tool achieved the highest accuracy (ACC 58.5% and AUC 0.62). The second best performance was achieved by the PROSSO II tool (ACC 58.0% and AUC 0.60). Surprisingly, modern tools based on deep learning techniques achieved an accuracy of around 52%.

These results show that the current tools are not yet very useful for the basic use case, i.e., protein solubility prediction. However, we wondered to what extent these tools can be used, at least for the task of protein prioritization in the context of the rational selection strategy. We, therefore, performed an experiment in which we sorted the test set samples for each tool according to the highest predicted score. We then retained only the top 10% of the highest scoring samples and verified how many of them were soluble. The best results were achieved by the SoluProt tool, which correctly identified 232 out of 310 proteins as soluble, which is a 49.7% improvement over the random selection, which would have correctly identified only 155 out of 310 samples. The other tools performed significantly worse in this test (ranging from -7.1% to 39.4%).

More detailed information about the developed SoluProt and the results of comparison against competing tools can be found in the original version of the manuscript in Appendix A.6, published in the Bioinformatics journal.

Although current tools cannot reliably determine protein solubility, it has been shown that they can be very useful for the prioritization task. We have therefore integrated our tool SoluProt as a key component of a pipeline for mining enzymes of interest from large databases (see next subsection).

Mining of soluble enzymes

We have developed EnzymeMiner, a tool that focuses on efficiently mining, filtering, annotating, and prioritizing enzymes of interest from large databases. It is the first tool of its kind to automate many steps that researchers have previously performed manually or using auxiliary scripts.

On its input, it expects the sequence of the enzyme of interest and information about the positions of essential residues necessary to maintain the enzyme's catalytic function. The EnzymeMiner tool then implements a three-step bioinformatics workflow: (i) homology search, (ii) essential residue-based filtering, and (iii) annotation of hits.

In the first step, the input sequence is used as a query for a PSI-BLAST [7] two-iteration search in the NCBI nr database [176]. The obtained hits are filtered in the second step using the input essential residue templates. Essential residues are checked using a global pairwise alignment with the template calculated by USEARCH [69] and a multiple sequence alignment calculated by Clustal Omega [185]. In the third step, the identified sequences are annotated using several databases and predictors: (i) transmembrane regions are predicted by TMHMM [111], (ii) Pfam domains are predicted by InterProScan [157], (iii) source organism annotation is extracted from the NCBI Taxonomy [73] and the NCBI BioProject database [17], (iv) sequence identities to original queries and resulting hits are calculated by USEARCH and (v) solubility is predicted by the solubility predictor SoluProt (described above).

The tool's output is an interactive table of found occurrences with rich annotation and the possibility of prioritization according to different criteria. This table is also complemented by graphical visualizations showing the network sequence similarity of the identified occurrences, which further facilitates their selection for experimental characterization purposes.

The EnzymeMiner workflow has been thoroughly experimentally validated using the model enzymes of haloalkane dehalogenases [206]. The sequence-based search identified 658 putative dehalogenases. The subsequent analysis prioritized and selected 20 candidate genes to explore their protein structural and functional diversity. The selected enzymes originated from genetically unrelated Bacteria, Eukarya, and, for the first time, also Archaea and showed novel catalytic properties and stabilities. The workflow helped to identify novel haloalkane dehalogenases, including (i) the most catalytically efficient enzyme ($k_{cat}/K_{0.5} = 96.8mM^{-1}s^{-1}$), (ii) the most thermostable enzyme showing a melting temperature of 71°C, (iii) three different cold-adapted enzymes active at near to 0°C, (iv) highly enantioselective enzymes, (v) enzymes with a wide range of optimal operational temperature from 20 to 70°C and an unusually broad pH range from 5.7-10, and (vi) biocatalysts degrading the warfare chemical yperite and various environmental pollutants. The sequence mining, annotation, and visualization steps from the workflow published by Vanacek et al. [206] were fully automated in the EnzymeMiner web server.

More detailed information about the developed EnzymeMiner tool can be found in the original version of the manuscript in Appendix A.7, published in the Nucleic Acid Research journal.

3.3 Rational protein design

3.3.1 State of the art

Identification of hotspots

As mentioned above, the basic idea of rational design is to engineer a small library of specific mutations (smart library) towards the desired change in the protein of interest. These hotspot mutations are selected based on a deep knowledge of the protein structure and function. The key role in the selection also depends on which property we wish to change. For example, catalytic properties such as activity, specificity, and stereoselectivity are often related to amino acid residues that mediate substrate binding, transition-state stabilization, or product release [60, 31]. Such residues can be identified using tools for predicting and analyzing enzyme-ligand interactions [222, 230, 118] or detecting binding pockets or access tunnels [182, 35, 232]. Strategies for improving protein stability include rigidifying flexible sites, cavity filling, tunnel engineering, consensus and ancestral mutation identification, or redesigning surface charges [30, 217, 229]. While hotspots for some of these strategies can be identified straightforwardly using a single computational tool [76], others require multi-step analyses or molecular modeling methods [18].

Although specific tools can be very effective, for example, in predicting mutations to increase protein stability, they may neglect other important properties such as activity [114, 184, 228]. Thus, it appears that when selecting suitable hotspots, it is essential to consider a number of different factors simultaneously, integrating the outputs of different tools and providing the user with a comprehensive view of the space of possible mutations, including their impact on protein structure and function.

To our knowledge, there is only one integral tool in this area so far, the HotSpot Wizard [22], developed within Loschmidt laboratories. This tool can combine the outputs of different bioinformatics analyses and automatically suggest sites suitable for mutagenesis.

However, the current version of the tool (2.0) has certain limitations, including the requirement for a 3D protein structure (mandatory input) and a large number of candidate mutations in the output that could be further filtered or prioritized.

Identification of mutations increasing protein stability

The application of proteins in various medical, biotechnological, and industrial applications often requires them to perform their function in an environment significantly different from the native one. A key factor is, therefore, the stability of the protein determining its applicability under harsh conditions such as extreme temperatures [12], acidic or basic pH, or unfavorable effects of organic solvents and proteases [151].

Stability is strongly connected with protein's conformation and can be qualified as the net balance of various intramolecular interactions and conformational entropy [84]. These interactions and forces can be strengthened or disrupted by introducing mutations into the protein. Therefore, many tools and methods have been developed to predict the effect of mutation on protein stability.

A large group of them consists of so-called energy-based methods that are based on molecular modeling of the physical interactions between the atoms in the tertiary structure of the protein. This group includes: Rosetta suite [103], ERIS software [227], Concoord/Poisson-Boltzmann surface area method [23], PopMuSiC method [59], DMutant [91] or HotMuSiC [155], CUPSAT [144], and FoldX suit [180].

The disadvantage of these tools is the high computational complexity and the requirement for knowledge of the protein's 3D structure. Therefore, alternative methods have emerged that attempt to identify suitable mutations based solely on the input sequence and the study of its evolution. Tools such as 3DM [112], VectorNTI [127], and EMBOSS [167] can be used for this purpose. This category also includes tools focused on ancestral sequence reconstruction, such as FastML [11], RAxML [194], Ancestors [64], HandAlign [216], and MrBayes [171] methods.

Machine learning-based methods have also been developed to reduce computational complexity. Their main advantage is that they do not require comprehensive knowledge of the physical and biochemical forces acting within a protein's tertiary structure. Predictions are therefore based exclusively on the available experimental data. This group includes tools such as: I-Mutant [39], EASE-MM [76], MuStab [202], and MuPro [49], PROTSRF [125], ProMaya [212], ELASPIC [220], and MAESTRO [116]. Unfortunately, independent studies [152, 103, 104] have shown that ML-based methods do not achieve the accuracy of energy-based methods, and their performance is often overestimated [154, 205]. This is mainly due to the limited experimental data necessary to train these tools.

The last group of tools is based on a combination of the above approaches, the so-called hybrid method. These methods use, for example, evolutionary analysis to identify conserved and correlated positions of a protein. Mutations at these positions are excluded from further analyses as they can seriously damage the structure or function of the protein. The space of candidate mutations is thus significantly reduced, decreasing the computational demands in subsequent energy-based calculations. This group includes, for example, the FRESCO protocol [218] and the PROSS tool [81].

In summary, hybrid methods represent the next step in predicting protein stability, as their robustness and complexity allow for constructing significantly more stable multiple-point mutants while maintaining reasonable computational demands. Unfortunately, currently, we are aware of only one server for the design of stable multiple-point mutants (PROSS) since FRESCO represents only a protocol description.

Research objectives

The aim of this work in rational design is to address two key topics:

- Eliminate the drawbacks of the current version of the HotSpot Wizard tool and extend it with 3D protein structure prediction and appropriate output mutation filtering mechanisms.
- Focus on identifying mutations leading to increased protein stability and develop a new tool based on a hybrid approach capable of predicting multiple-point mutants.

3.3.2 Research summary

HotSpot Wizard 3.0

We have created a new version of the HotSpot Wizard tool, developed within the Loschmidt laboratories. The basic idea of this tool is to identify positions in proteins called hotspots that will be suitable for mutagenesis. For this purpose, the tool uses four strategies:

- *Identification of functional hotspots* – the tool searches highly mutable residues located in the catalytic pockets or tunnels connecting these pockets with the bulk solvent. Residues located close to the active site have been identified as good mutagenesis targets for engineering [163, 161, 137]. On the other hand, catalytic residues are excluded from this list to avoid damaging the primary function of the protein.
- *Identification of stability hotspots (based on structural flexibility)* – based on the calculation of B-factors, positions suitable for stabilization of flexible regions of the protein are identified. The rationale for targeting these flexible residues is that they have relatively few contacts with neighbors, so their substitution can produce more interactions [162, 41, 97].
- *Identification of stability hotspots (based on conservation)* – implements majority and frequency ratio approaches, both of which suggest mutations at positions where the wild-type amino acid differs from the most prevalent amino acid (i.e., the consensus residue) at a given position in the multiple sequence alignment. The assumption that the most common amino acid is likely to be stabilizing has proven to be very successful at creating more stable proteins [199, 146, 9, 119].
- *Identification of correlated hotspots* – involves searching for coordinated changes of the amino acids at two separate positions within the protein-co-evolving residues. These correlated positions are subsequently removed from the list of hotspots, as they may break key links that ensure protein stability or the function of interfaces used for interaction between proteins.

The resulting pipeline of twenty integrated tools and three databases represents a unique one-stop solution that makes library design accessible even to users with no prior knowledge of bioinformatics.

When designing and implementing the new version (3.0), we addressed two major drawbacks of the previous version: (i) the requirement for the protein tertiary structure as essential input and (ii) a large number of candidate hotspots in the output of the tool. HotSpot Wizard 3.0 (HSW3) dramatically enhanced usability by overcoming these key limitations.

If only a protein sequence is used as input, the HSW3 first verifies the existence of a 3D structure for that sequence. It searches the RCSB Protein Data Bank [26] and, if unsuccessful, the Protein Model Portal [86], which collates models of protein structures from different resources. If no existing model can be found for the protein, the tool proceeds to homology modeling. For this purpose, the user has a choice between a pair of tools: Modeller [215] (faster but less accurate) and I-Tasser [225] (more accurate but slower).

It is essential to verify the quality of the created model before using it. Therefore, HSW3 provides a robust quality assessment of the protein structure using three well-established tools: PROCHECK [117], MolProbity [47], and WHAT CHECK [90].

In addition to the prediction of the 3D protein structure, the HSW3 was also enriched to evaluate the proposed mutations (outputs from the four main strategies) in terms of their effect on protein stability. Wild-type and mutant structures were evaluated using Rosetta [103] software. Finally, FoldX [180] tool was used for repairing protein structure by filling in the missing atoms and patching the structure. In summary, this step allowed us to filter out a huge number of mutations that would destabilize the protein and thus make experimental characterization more expensive.

The final version of the HSW3 web server has been thoroughly validated and tested. The reliability of the Rosetta protocol was benchmarked against experimental stability data previously collected for multiple-point mutants in the Loschmidt laboratory [18] as well as 1,573 single-point mutants available in the ProTherm database [113] and HotMuSiC dataset [155]. These tests confirmed a significant correlation between half-lives and calculated changes in free energy G, as well as the ability of the Rosetta protocol to classify stabilizing and destabilizing mutations correctly. The quality of HSW3 predictions was further validated by saturation mutagenesis at the hotspot position L177 located at the tunnel mouth of the haloalkane dehalogenase LinB [42]. Theoretical predictions correctly identified the variant L177W, which was also found to be the most stable experimentally.

More detailed information about the developed HotSpot Wizard 3.0 tool and its validation can be found in the original version of the manuscript in Appendix A.8, published in the Nucleic Acid Research journal.

FireProt

We have created a new tool FireProt for the automated design of thermostable proteins. FireProt combines energy- and evolution-based approaches to design thermostable multiple-point mutants. This combination appears to be very useful since phylogenetic analysis enables the identification of the mutations stabilized by entropy, which cannot be predicted by force field calculations [20].

First, the FireProt tool uses evolution-based approaches to identify conserved and correlated positions. These are excluded from further analysis as it was observed that functional and structural constraints in proteins generally lead to the conservation of amino acid residues [24, 34, 54, 92]. Similarly, correlated residues ordinarily help to maintain protein function, folding, or stability [80, 139, 201]. Mutations conducted on these positions are therefore considered unsafe by the current FireProt strategy.

The remaining positions are subjected to an energy-based approach and saturation mutagenesis by using the FoldX tool. Mutations with predicted $\Delta\Delta G$ over a given threshold are steered away, and the rest is forwarded to Rosetta calculations. Finally, the mutations predicted by Rosetta as strongly stabilizing are tagged as potential candidates for the design of the multiple-point mutants.

Because of potentially antagonistic effects between individual mutations, we cannot combine individual mutations blindly. To avoid possible clashes, the FireProt strategy tries to minimize antagonistic effects by utilizing Rosetta. In the first step, all pairs of single-point mutations within the range of 10 Å are evaluated. Once a change in free energy is obtained for all residue pairs, FireProt starts to introduce them into the multiple-point mutant in the order based on their predicted stability, excluding the mutations that are colliding with already included mutations. The algorithm stops once no mutations are left or the analyzed pair's stabilizing effect drops below a defined threshold.

In summary, FireProt integrates sixteen computational tools and utilizes sequence and structural information. It represents a unique integrated solution that makes the design of thermostable proteins accessible even to users with no prior knowledge of bioinformatics.

The FireProt protocol was experimentally verified with three proteins (haloalkane dehalogenase DhaA, hexachlorocyclohexane dehydrochlorinase LinA, and fibroblast growth factor 2). It provided higher stabilization of proteins from 15 to 25°C compared to wild-type. Additionally, FireProt predictions of eight multiple-point mutants were validated using the PROSS tool. FireProt and PROSS showed similar predictive power, correctly identifying 29 and 20 potentially stabilizing positions.

More detailed information about the FireProt tool and its validation can be found in the original version of the manuscript in Appendix A.9, published in the Nucleic Acid Research journal.

3.4 List of publications

Publication I

Title	SoluProt: prediction of soluble protein expression in Escherichia coli
Authors	HON Jiří, MARUŠIAK Martin, MARTÍNEK Tomáš, KUNKA Antonín, ZENDULKA Jaroslav, BEDNÁŘ David, and DAMBORSKÝ Jiří
Abstract	**Motivation:** Poor protein solubility hinders the production of many therapeutic and industrially useful proteins. Experimental efforts to increase solubility are plagued by low success rates and often reduce biological activity. Computational prediction of protein expressibility and solubility in Escherichia coli using only sequence information could reduce the cost of experimental studies by enabling prioritization of highly soluble proteins. **Results:** A new tool for sequence-based prediction of soluble protein expression in E.coli, SoluProt, was created using the gradient boosting machine technique with the TargetTrack database as a training set. When evaluated against a balanced independent test set derived from the NESG database, SoluProt's accuracy of 58.5% and AUC of 0.62 exceeded those of a suite of alternative solubility prediction tools. There is also evidence that it could significantly increase the success rate of experimental protein studies. SoluProt is freely available as a standalone program and a user-friendly webserver at `https://loschmidt.chemi.muni.cz/soluprot/`. **Availability:**`https://loschmidt.chemi.muni.cz/soluprot/`.
Journal	Bioinformatics, vol. 37, num. 1, 2021 Journal impact factor: 6.931, Q1
Citations	15 (WoS without self-citations)
Author's contribution	Studying state-of-the-art, consulting on algorithm design and implementation including preparation of datasets.
Manuscript	Appendix A.6

Publication II

Title	EnzymeMiner: automated mining of soluble enzymes with diverse structures, catalytic properties and stabilities
Authors	HON Jiří, BORKO Simeon, ŠTOURAČ Jan, PROKOP Zbyněk, BEDNÁŘ David, ZENDULKA Jaroslav, MARTÍNEK Tomáš, and DAMBORSKÝ Jiří
Abstract	Millions of protein sequences are being discovered at an incredible pace, representing an inexhaustible source of biocatalysts. Despite genomic databases growing exponentially, classical biochemical characterization techniques are time-demanding, cost-ineffective and low-throughput. Therefore, computational methods are being developed to explore the unmapped sequence space efficiently. Selection of putative enzymes for biochemical characterization based on rational and robust analysis of all available sequences remains an unsolved problem. To address this challenge, we have developed EnzymeMiner—a web server for automated screening and annotation of diverse family members that enables selection of hits for wet-lab experiments. EnzymeMiner prioritizes sequences that are more likely to preserve the catalytic activity and are heterologously expressible in a soluble form in Escherichia coli. The solubility prediction employs the in-house SoluProt predictor developed using machine learning. EnzymeMiner reduces the time devoted to data gathering, multistep analysis, sequence prioritization and selection from days to hours. The successful use case for the haloalkane dehalogenase family is described in a comprehensive tutorial available on the EnzymeMiner web page. EnzymeMiner is a universal tool applicable to any enzyme family that provides an interactive and easy-to-use web interface freely available at `https://loschmidt.chemi.muni.cz/enzymeminer/`.
Journal	Nucleic Acids Research, vol. 48, num. 1, 2020 Journal impact factor: 16.971, Q1
Citations	16 (WoS without self-citations)
Author's contribution	Implementation of selected modules, consultation on the design and implementation of the tool.
Manuscript	Appendix A.7

Publication III

Title	HotSpot Wizard 3.0: Web Server for Automated Design of Mutations and Smart Libraries based on Sequence Input Information
Authors	SUMBALOVÁ Lenka, ŠTOURAČ Jan, MARTÍNEK Tomáš, BEDNÁŘ David, and DAMBORSKÝ Jiří
Abstract	HotSpot Wizard is a web server used for the automated identification of hotspots in semi-rational protein design to give improved protein stability, catalytic activity, substrate specificity and enantioselectivity. Since there are three orders of magnitude fewer protein structures than sequences in bioinformatic databases, the major limitation to the usability of previous versions was the requirement for the protein structure to be a compulsory input for the calculation. HotSpot Wizard 3.0 now accepts the protein sequence as input data. The protein structure for the query sequence is obtained either from eight repositories of homology models or is modeled using Modeller and I-Tasser. The quality of the models is then evaluated using three quality assessment tools—WHAT CHECK, PROCHECK and Mol-Probity. During follow-up analyses, the system automatically warns the users whenever they attempt to redesign poorly predicted parts of their homology models. The second main limitation of HotSpot Wizard's predictions is that it identifies suitable positions for mutagenesis, but does not provide any reliable advice on particular substitutions. A new module for the estimation of thermodynamic stabilities using the Rosetta and FoldX suites has been introduced which prevents destabilizing mutations among pre-selected variants entering experimental testing. HotSpot Wizard is freely available at `http://loschmidt.chemi.muni.cz/hotspotwizard`.
Journal	Nucleic Acids Research, vol. 46, num. 1, 2018 Journal impact factor: 11.147, Q1
Citations	85 (WoS without self-citations)
Author's contribution	Consultation on the design and implementation of the tool, especially the module for 3D protein structure prediction.
Manuscript	Appendix A.8

Publication IV

Title	FireProt: web server for automated design of thermostable proteins
Authors	MUSIL Miloš, ŠTOURAČ Jan, BENDL Jaroslav, BREZOVSKÝ Jan, PROKOP Zbyněk, ZENDULKA Jaroslav, MARTÍNEK Tomáš, BEDNÁŘ David, and DAMBORSKÝ Jiří
Abstract	There is a continuous interest in increasing protein stability to enhance their usability in numerous biomedical and biotechnological applications. A number of in silico tools for the prediction of the effect of mutations on protein stability have been developed recently. However, only single-point mutations with a small effect on protein stability are typically predicted with the existing tools and have to be followed by laborious protein expression, purification, and characterization. Here, we present FireProt, a web server for the automated design of multiple-point thermostable mutant proteins that combines structural and evolutionary information in its calculation core. FireProt utilizes sixteen tools and three protein engineering strategies for making reliable protein designs. The server is complemented with an interactive, easy-to-use interface that allows users to directly analyze and optionally modify designed thermostable mutants. FireProt is freely available at `http://loschmidt.chemi.muni.cz/fireprot`.
Journal	Nucleic Acids Research, vol. 45, num. 1, 2017 Journal impact factor: 11.561, Q1
Citations	55 (WoS without self-citations)
Author's contribution	Consultation on the design and implementation of the tool.
Manuscript	Appendix A.9

3.5 Conclusions

In the field of protein engineering, we focused on improving existing approaches in the areas of rational selection and rational design.

As part of the rational selection strategy, we have developed the EnzymeMiner tool, which can efficiently mine information about enzymes of interest from large databases and perform filtering and prioritization concerning various criteria. We identified protein solubility as one of the key criteria, as it fundamentally affects subsequent experimental protein characterization. Therefore, we also focused on developing a new solubility prediction tool. The developed tool, SoluProt, uses machine learning techniques and achieves the best prediction accuracy and prioritization capability compared to existing approaches. It has, therefore, also been integrated as part of the EnzymeMiner tool.

As part of the rational design strategy, we have contributed to developing a new version of the HotSpot Wizard (3.0) that allows users to identify suitable positions for mutagenesis. This new version has been extended with the ability to predict the 3D structure of a protein and to evaluate the effect of mutations on protein stability. These new features have significantly expanded the possibilities of using this tool while eliminating many potential mutations leading to protein destabilization.

For the use of proteins in real applications in medicine or industry, it is often required that the protein of interest can work even under harsh conditions (e.g., higher temperature, pH, etc.). Therefore, we also focused on developing a new tool for the automated design of thermostable proteins. The developed tool FireProt efficiently combines energy-based and evolutionary-based approaches to design thermostable multiple-point mutants. This tool's applicability has been demonstrated experimentally and in comparison with competing tools.

3.5.1 Future work

At present, we are still far from being able to consider the developed approaches and tools as final.

EnzymeMiner

The developed tool for mining enzymes of interest could be integrated with other types of databases. For example, the MGnify metagenomics database [136] of more than 267 million protein sequences can be a rich source of information. The MGnify database contains proteins from organisms living deep in the ocean in hot springs or digestion systems. Proteins from such organisms may have interesting properties for biotechnological applications.

If we could build a 3D protein structure for the hits found (e.g., through homology modeling), it would be possible to add support for various additional annotation and efficient filtering techniques to the tool. For example, information about the size of the catalytic pocket or the structural properties of the tunnels going to the active site could be beneficial.

Given the ever-increasing data in current biological databases, it is advisable to repeat the mining process periodically and inform users about new candidate sequences. EnzymeMiner should therefore be extended with a so-called incremental mining mode, which would focus only on newly added data and thus be undoubtedly more efficient in terms of computational effort.

SoluProt

Solubility prediction tools are moving toward determining not only the global solubility of a protein but also the solubility profile (solubility score for individual residues). This ability, in turn, will also allow the prediction of the effect of mutation on solubility and, thus, the development of new approaches for increasing protein solubility by appropriate mutations.

It also shows that knowledge of the 3D structure of a protein can significantly help in predicting the effect of a mutation on solubility. Thus, in the future, it would be helpful to extend the tool to predict the 3D structure of the protein of interest (e.g., through homology modeling) and enrich the predictor with several structural features.

Finally, novel experimental data for protein solubility change upon a single-point mutation are emerging rapidly thanks to the advent of deep mutational scanning technology [108]. The data usually contain thousands of samples covering nearly all possible point mutations in a selected protein, making them well suited for understanding the fundamental mechanisms influencing protein solubility.

HotSpot Wizard

The current version of the tool could be extended, for example, by a more detailed analysis of correlated mutations. While the current version eliminates them for safety reasons, the new version could allow combinations that would not disrupt the structure and function of the protein.

The prediction of positions suitable for mutagenesis represents an area that can be continuously enriched with new information on protein structure and function. Therefore, it is essential to equip the HSW3 with as general an interface as possible to integrate other tools and analyses in the future.

FireProt

Similar to the HotSpot Wizard, it would be helpful to perform a more detailed analysis of correlated mutations (especially contacts) and use them to increase protein stability. A challenge for current tools is a more detailed understanding of the importance of charges on the protein surface, as mutations in these regions have been shown to affect stability significantly.

Finally, with the advent of new technologies and more experimental data, it would be useful to reopen the idea of using machine learning-based tools that have the potential to model very complex relationships within a protein at a reasonable computational cost.

3.6 Research impact

All developed tools are quite complex and often integrate many external tools. Moreover, some of the implemented analyses are computationally intensive, especially where they involve energy-based calculations or 3D protein structure prediction. Distributing these tools as open-source is usually not efficient, as end-users are often unable to install and operate such complex systems. A variant in the form of a web-based interface further connected to the computational core of the application running in a grid or cloud environment seems to be preferable. All four tools presented in this section were created in this way. Users access them through a web interface, and their input requests are sorted into a queue of jobs, executed sequentially using the Metacenter grid infrastructure. Once a job is processed, a notification is sent to the user with a link to the results.

In the following Table 3.1, you can see the statistics of the number of visits to the website for each of the tools created, including the number of completed jobs. This information is supplemented by the number of citations at the end. From these statistics, it can be concluded that the tools created are widely used and cited by the scientific community.

Tool	**Number of visitors**	**Number of jobs**	**Number of citations**
EnzymeMiner `https://loschmidt.chemi.muni.cz/enzymeminer/`	13,471	2,799	16
SoluProt `https://loschmidt.chemi.muni.cz/soluprot/`	13,281	23,484	15
HotSpot Wizard `https://loschmidt.chemi.muni.cz/hotspotwizard/`	71,787	38,068	85
FireProt `https://loschmidt.chemi.muni.cz/fireprotweb/`	30,965	5,582	55

Table 3.1: Statistics on the use of developed tools

Chapter 4

Repetitive DNA sequences

4.1 Introduction

Repetitive sequences are pieces of DNA that occur in an excessive number of copies in genomes. They appear to be part of the genomes of most living organisms, from simple bacteria to complex eukaryotes, including plants and animals. For example, studies report that repetitive sequences make up more than two-thirds of the human genome [110]. Their abundance is even higher in plants, e.g., 80% of the maize genome [178]. Initially, these sequnces were considered as „junk“ DNA, but more and more studies have demonstrated their importance in many biological processes [147, 149, 50, 133, 123].

Repetitive sequences fall into two main categories: tandem repeats and interspersed repeats, often referred to as transposable elements.

4.1.1 Tandem repeats

Tandem repeats (TR) are defined as repetitive pieces of DNA placed side by side in a large number of copies. The core of the repeated sequence is referred to as a monomer. According to the length of this monomer, tandem repeats are divided into: (i) microsatellites with monomer length < 9 nucleotides, (ii) minisatellites with monomer length between 10 and 100 bp, and (iii) satellite DNA (satDNA) having monomers longer than 100 bp. Based on sequence similarity, we then classify monomers into different families. For example, there are 12 satDNA families in Hippophae rhamnoides [156], 62 families in Locusta migratoria [173], or 9 families within the human genome [134].

Although TRs were initially considered to be non-functional DNA, at present, we know they have many functions in the genome. TRs are involved in chromosome organization, telomere elongation control, transcriptional response during stress, or the modulation of gene expression [147, 149]. They could influence the adaptability of a host genome and sex chromosome evolution [46].

4.1.2 Transposable elements

Transposable elements (TEs), also called jumping genes, were firstly discovered in the 1940s by geneticist Barbara McClintock [16]. They can move or even copy themselves from one genomic location to another, resulting in their rapid amplification in the genome. TEs can occur in hundreds or even thousands of copies.

TEs are divided into two major classes according to whether they transpose via an RNA intermediate, Class I - retrotransposons, or a DNA intermediate, Class II - DNA transposons. Both of these groups are further subdivided into subclasses and families of TEs. For example, Class I - Retransposons include LTR retrotransposons (LTR), Dictyostelium Intermediate Repeat Sequences (DIRS), Penelope-like elements (PLE), Long Interspersed Nuclear Elements (LINE), and Short Interspersed Nuclear Elements (SINE). Similarly, Class II - DNA transposons are further subdivided into subclasses: TIR, Crypton, Helitron, and Maverick.

The individual classes and subclasses of TEs are specific in their structure. Typically, they contain a group of protein-coding genes essential for their transmissions, such as reverse transcriptase (RT), transposase (TPase), integrase (INT), or tyrosine recombinase (YR). The structure of TEs can also be enriched with a number of other signatures, such as long-terminal repeats (LTR), terminal inverted repeats (TIR), direct repeats (DR), Poly(A), or A- or AT-rich regions. Examples of TEs structures can be seen in Figure 4.1.

Recent studies have revealed that TEs are involved in several biological processes. They can regulate genes [50, 130], increase genetic variation, influence genome size [133, 48], play an essential role in chromosomal rearrangements [123], or be crucial players in genome evolution [25, 32].

4.1.3 Study of repetitive sequences

With the advent of high-throughput sequencing technologies, research on repetitive sequences and their representation in genomes has also been significantly expanded.

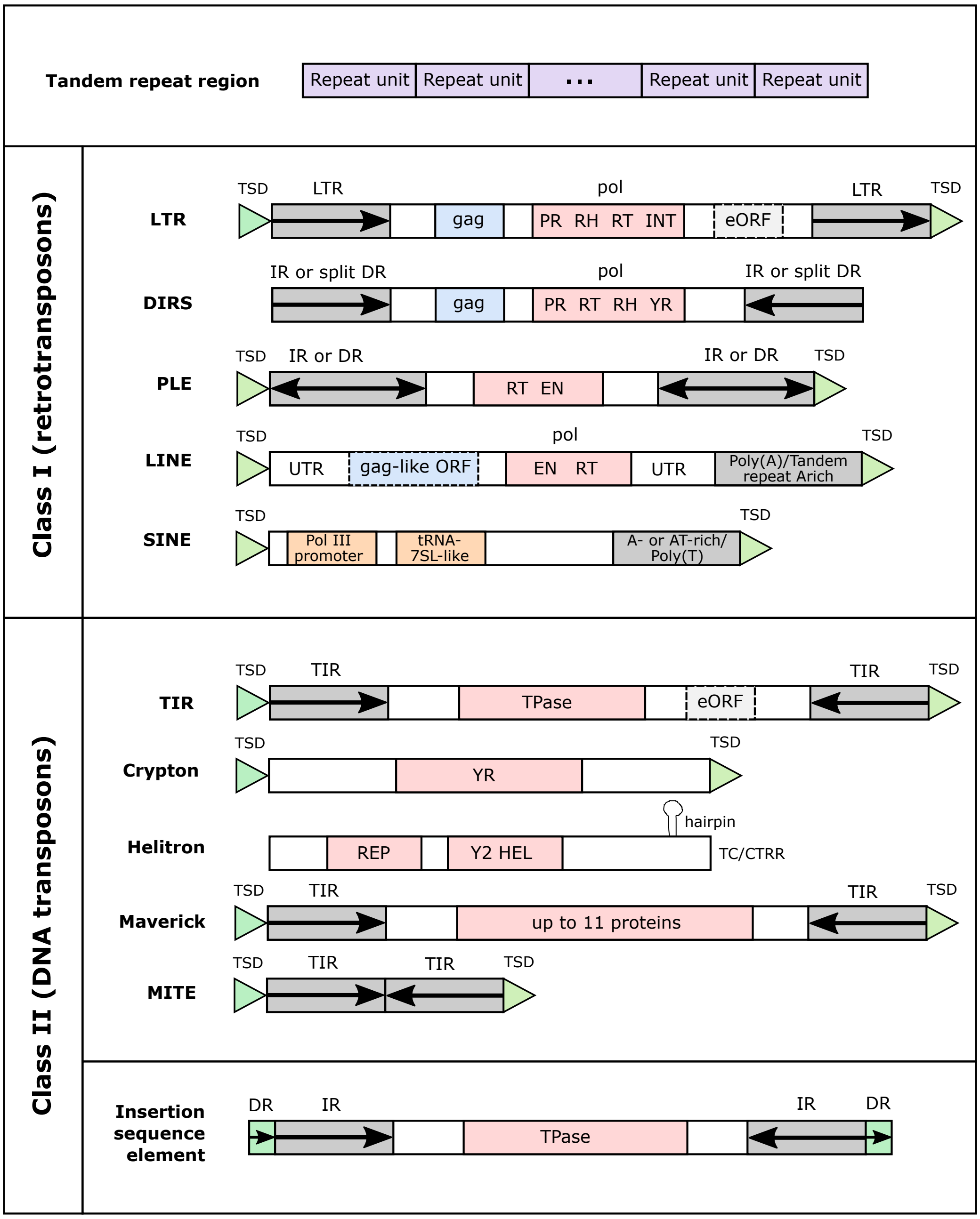
One of the basic techniques is the study of repeats in the already assembled genome, i.e., the complete sequence divided into individual chromosomes. However, obtaining an assembled genome is not easy. The organism's genome must first be sequenced with considerable coverage. Short sequencing reads are subsequently linked into longer contigs, supercontigs, up to the resulting chromosomes. Interestingly, the process of genome assembly is significantly complicated by the presence of a large number of repetitive sequences. Linking repetitive sequences into unique contigs is often complex and ambiguous.

This issue can be partially improved either by higher genome coverage (increasing the cost) or by using third-generation sequencing technologies that produce long reads (up to several tens of thousands of bp) that can bridge the repeat and correctly link the corresponding contigs. The second method is, in principle, more efficient, but for now, third-generation sequencing technologies are still under development and struggling with high error rates.

As an alternative method, so-called low-pass sequencing is applied, where the genome of the organism of interest is sequenced with low coverage (e.g., around 5-10%). Sequencing reads of such low coverage cannot be used to assemble the genome (or even a part of it). However, this tiny amount of data will contain a sufficient number of copies of repetitive sequences present in several thousand copies within the original genome. The low-pass sequencing technique thus represents a good compromise between the cost of sequencing and the amount of information obtained.

Based on the input data type, methods for searching and analyzing repetitive sequences are divided into assembly-based and assembly-free approaches.

Figure 4.1: Major classes of repetitive elements and examples of their families and structural features. DR - direct repeat, EN - endonuclease, eORF - extra open reading frame, gag - gag gene, INT - integrase, IR - inverted repeat, LTR - long terminal repeat, ORF - open reading frame, pol - pol gene, PR - protease, RH - RNase H, RT - reverse transcriptase, UTR - untranslated region, TSD - target site duplication, YR - tyrosine recombinase, HEL - helicase domain, REP - replication initiator motif, TIR - terminal inverted repeat, TPase - transposase gene, TSD - target site duplication, YR - tyrosine recombinase, Y2 - Y2-type tyrosine recombinase.



4.2 State of the art

4.2.1 Assembly-based approaches

Several approaches are used to search for repetitive sequences in assembled genomes. The first category consists of library-based methods. Their basic idea is to explore the input sequence and compare it with a database of known repeats, such as Repbase [14]. The comparison itself is performed based on sequence similarity, whereby tools like BLAST [8] or HMMER [68] are used. The range of repeats found strongly depends on the quality of the database. Unfortunately, these methods cannot find new (previously undiscovered) repeats by design.

Signature-based methods represent the second group. The goal of these methods is to search for signatures specific to TEs. For example, LTR retrotransposons are terminated with long terminal repeats at their 5' and 3' ends. In contrast, the TIR, Maverick, and MITE families have inverted repeat regions. These specific signatures can be searched for regardless of knowledge of the full-length TE sequence. Therefore, compared to library-based methods, these approaches can also find new families/subfamilies of TEs. On the other hand, they often suffer from a large number of false positives that need to be further analyzed and filtered. Tools in this category include: LTR STRUCT [131], LTR par [98], and detectIR [226].

A separate group consists of so-called de novo methods, which aim to find all repetitive sequences regardless of prior knowledge of their sequence or structure. This goal can be achieved, for example, by finding all local similarities of an input sequence to itself through tools such as Repeat Pattern Toolkit [4], RECON [15], and PILER [70]. For large genomes, however, this approach is very computationally intensive. It is therefore being replaced by more efficient k-mer counting-based methods such as Reputer [115], RepeatScout [153], and RepLoc [74]. The drawbacks of these methods include the difficulty in detecting repeats with low copy numbers and the need for appropriate adjustment of the k-mer length to achieve the desired sensitivity.

The last group is represented by hybrid methods that combine some of the above approaches. Probably the most common is the combination of library-based and signature-based techniques. While signature-based search is used, for example, to detect terminal LTR regions, the library-based approach verify the presence of gag or pol coding regions. Such a combination can detect new LTR transposon subfamilies and reduce many false positives typical for signature-based methods. Tools in this category include, for example, LTR finder [224], LTRdigest [196], and TIRfinder [79].

Please note that the selection of a particular method or combination of approaches always depends on the structure of the TE being searched. For example, there are TEs without any signatures such as Crypton and Helitron, or their signatures may be very short (in the range of units of bases). In these cases, signature-based approaches cannot be used.

4.2.2 Assembly-free approaches

Most of the tools for searching, reconstructing, and quantifying repeats directly from sequencing data are based on data from 2nd generation sequencers. This is mainly due to their availability and low error rate. They expect single or pair-end reads of about 200-300bp as inputs. On the other hand, the examined repeats are usually longer (units to tens of thousands of bp), and therefore these tools have to deal with their assembly. With the increasing availability of data from 3rd generation sequencers (long reads with higher error rates), so-called hybrid approaches combining the 2nd and 3rd generation data are starting to emerge. While the 3rd generation long reads are used to identify the underlying scaffold structure of the repeat, the 2nd generation data are additionally mapped to refine their content. Generally, tools for reconstructing repeats from sequencing data fall into three basic groups.

The first group consists of approaches based on the k-mer counting technique, similar to the assembly-based methods. The basic idea of these tools is to split reads into individual k-mers, identify the most numerous k-mers, and gradually concatenate them up to the level of reconstructed repetition. This group includes tools such as: ReAS [122], RepARK [109], REPdenovo [51], and DLR [126]. The disadvantage of these approaches is the difficulty in detecting repeats with low copy numbers and reconstructing regions with lower sequence similarity, for example, evolutionarily distant families or older TEs with higher numbers of accumulated mutations.

The second group of methods uses de Bruijn graphs to reconstruct repetitive sequences. The de Bruijn graph is a directed multigraph consisting of vertices and a multiset of directed edges. It is constructed using the unique k-mers that occur in the input data (sequencing reads). k-1-mers are added to the graph as nodes and k-mers as edges. To assemble the original genome, every edge in the graph is visited exactly once, representing an Eulerian path problem that can be resolved in a linear time. These types of graphs have therefore been successfully used for DNA sequence assembly and implemented in tools such as Trinity [83], ABySS [188], and Velvet [231]. Similar principles have been adopted by repeat reconstruction tools such as Tedna [236], dnaPipeTE [82], and MixTaR [75]. Unfortunately, the variability of repetitive sequences combined with the error rate of sequencing reads causes the frequent occurrence of branching structures in the form of tips („dead-ends“) and bubbles within de Bruijn graphs. These branching structures make the constructed graph very large, leading to high memory usage and increased computational demand.

The third group of methods also uses graphs, but in a different way. The graph nodes represent individual reads, and edges connect reads that achieve a certain level of sequence similarity. Before constructing such a graph, an all-against-all pairwise alignment of the input reads is performed, representing one of the most computationally demanding steps. Subsequently, the graph is scanned for connected components, groups of mutually connected vertices representing repetitive sequences. This group includes tools such as: RepeatExplorer [140], Transposome [195], and RepLong [85].

In summary, graph-based clustering approaches can deal better with the variability of repetitive sequences compared to k-mer or de Bruijn-graph-based approaches as the desired similarity and overlap length thresholds can be set. Another significant advantage is that the input reads do not have to be split into k-mers, thus their continuity is not lost. The main drawbacks of the graph-based clustering approach are (i) chimeric clusters and (ii) splitting one repeat family into multiple clusters. Both happen due to divergence of repeat and different conservation levels within the repeat families. For these reasons, it is recommended to manually inspect the output of these tools and refine the created clusters.

4.2.3 Research objectives

This work focuses on improving existing methods in both assembly-based and assembly-free approaches. One of the main goals is to develop a new tool for detecting Insertion Sequence elements in assembled prokaryotic genomes. Other goals include the development of a new technique for the reconstruction and analysis of satellite DNA directly from sequencing data.

4.3 Research summary

4.3.1 digIS - novel approach for detection of distant Insertion Sequence elements

Insertion Sequence elements (ISE) are TEs widespread in prokaryotic genomes. They play an essential role in genome evolution, structure, and host-genome adaptability, including modulation of gene expression or antimicrobial resistance [187, 207]. Their body typically encodes a protein that catalyzes the transposition (TPase), flanked by short IRs and DRs (see Figure 4.1). Up to now, 29 IS families have been identified [187].

At present, there are several tools available for the detection of IS elements in prokaryotic genomes. Some of them are designed for searching in raw sequenced data (ISQuest [29], ISMapper [87], ISseeker [3], and panISa [204]), and the others require assembled sequences (IScan [211], ISsaga [208], OASIS [170], ISEScan [221], and TnpPred [166]). Since signatures in the form of IR and DR are very weak and in some families not present at all, most tools utilize a library-based approach which is dependent on a source of known IS elements, usually ISfinder database [186].

Unfortunately, the developed tools are either very conservative and accept only sequences that are very close to known ISEs, or they are benevolent and report even fragments of ISEs in their outputs, including many false positives. Therefore, this work aimed to develop a new approach that would search not only for known ISEs but also for members of putative novel families while reporting the lowest number of false positives.

In developing the digIS tool, we designed a novel approach that first targets the TPase catalytic domain, representing the most conserved part of the ISE. Based on known ISE sequences from the ISfinder database, we built and manually refined profile HMM models of the catalytic domains for all available families. Using the created models and the HMMER tool [135], we then searched the input genome sequence translated into all six frames.

The occurrences found are treated as seeds, which are then filtered, merged, and expanded based on similarity to sequences of known ISEs in the ISfinder database. If a GenBank annotation is available for the input sequence, the tool classifies the found occurrences into three categories that help the user better distinguish the quality of the hits. The search and classification results are finally saved in a GFF3 format file.

We compared the performance of digIS with competing tools on manually annotated datasets from the ISbrowser database [106] and automatically annotated genomes from the NCBI database [176]. The results of the comparisons demonstrated that the developed tool could identify not only known ISEs but also putative novel elements. Compared to tools that also detect fragments, the digIS tool reports significantly fewer false positives.

More detailed information about the digIS tool can be found in the original version of the manuscript in Appendix A.10, published in the BMC Bioinformatics journal.

4.3.2 Novel approach for detailed analysis of satellite DNA

While studying repetitive regions of the seabuckthorn (Hippophae rhamnoides) genome, we developed a new approach for satellite DNA analysis. This approach is based directly on sequencing data and extends the comprehensive analysis of repetitive regions of the RepeatExplorer tool [140]. The core of the proposed method operates in three steps:

- *Detection of satellite monomers* – Contigs of selected clusters are extracted from RepeatExplorer output. For each contig, the monomer length is estimated from distances between the identical k-mers in the contig. Finally, the monomer sequence is extracted from the most covered region of the contig.
- *Estimation of satellite family composition and their annotation* – To estimate the composition of satellite families, extracted monomer sequences are clustered using the UPGMA method [192]. The resulting dendrogram is cut to define the individual satellite families and visualized using igraph library[1]. Then, monomers are annotated by querying them against the nt/nr nucleotide collection and PlantSat database [128] using BLASTN [8]. To estimate the diversity within each satellite family, reads belonging to the family are mapped to the representative monomer using the BWA-MEM aligner [121], and a sequence logo is generated by WebLogo [55].
- *Visualization of satellite families' homogeneity* – Reads of each satellite family are merged and sampled randomly to decrease the computational demands for highly abundant families. Sequence similarity of these reads is estimated by all-against-all alignment performed by MegaBLAST [38]. Only pairs of reads that meet the specific thresholds (70% sequence identity over at least 55% sequence length) are used for graph construction and visualization. A relative abundance of male and female reads in each family is estimated. Such information can be useful to determine the chromosomal location, whether the satellite family is present on sex chromosomes or autosomes.

By applying the approach described above, we identified 12 satellite families in the seabuckthorn (Hippophae rhamnoides) genome, including Y-specific, X-accumulated, and sex-chromosome-accumulated satellite families. The discovery of the Y-specific satellite helped to show that seabuckthorn has small Y and large X chromosomes since it was previously thought to be exactly the opposite [72].

More details of the approach developed for satellite DNA analysis can be found in the original version of the manuscript in Appendix A.11, published in Genome Biology and Evolution journal.

4.3.3 Additional studies

In addition to developing specific tools or extending existing pipelines, we also participated in a study investigating the relationship between transposons in the human genome and secondary structures, specifically quadruplexes.

Positions of repetitive sequences in the human genome were collected using UCSC Table Browser data [100] (Repeat Masker track [101]) and extended with 200 bp flanking regions. The collected sequences were scanned for the occurrence of the typical quadruplex pattern GGG.{1,7}GGG.{1,7}GGG.{1,7}GGG[2] on both strands.

In subsequent analyses, we verified the occurrence of potential quadruplex-forming sequences (PQS) within and around the four most abundant families of TEs in the human

[1] `https://igraph.org/r/`

[2] Please note that our quadruplex search tool, pqsfinder (2017), was not yet available at the time of the study (2014).

genome (LINE-1, HERV, SVA, and ALU). We also extended the analysis to compare differences between TEs and PQSs on the X and Y chromosomes, including an analysis of the occurrence of PQSs in the vicinity of TEs of different ages. Finally, we selected 12 PQS sequences found in the vicinity of TEs and performed in vitro experimental evaluation using circular dichroism measurements and gel electrophoresis.

In summary, the study results suggest that the activity of transposable elements, especially LINE-1 and SVA elements, contributes toward genome-wide quadruplex distribution in humans. Conservation of quadruplexes at specific positions implies their function either in the life cycle of transposable elements or host genome maintenance, or both. All tested PQSs could form quadruplex structures in vitro, albeit with differing willingness, strand orientation, and molecularity. LINE-1 and SVA families displayed an age-dependent pattern with younger elements containing a higher number of more stable quadruplexes.

More detailed information on the results of this study can be found in the original version of the manuscript in Appendix A.12, published in the BMC Genomics journal.

4.3.4 List of publications

Publication I

Title	digIS: Towards detecting distant and putative novel insertion sequences in prokaryotic genomes
Authors	PUTEROVÁ Janka and MARTÍNEK Tomáš
Abstract	**Background:** The insertion sequence elements (IS elements) represent the smallest and the most abundant mobile elements in prokaryotic genomes. It has been shown that they play a significant role in genome organization and evolution. To better understand their function in the host genome, it is desirable to have an effective detection and annotation tool. This need becomes even more crucial when considering rapid growing genomic and metagenomic data. The existing tools for IS elements detection and annotation are usually based on comparing sequence similarity with a database of known IS families. Thus, they have limited ability to discover distant and putative novel IS elements. **Results:** In this paper, we present digIS, a software tool based on profile hidden Markov models assembled from catalytic domains of transposases. It shows a very good performance in detecting known IS elements when tested on datasets with manually curated annotation. The main contribution of digIS is in its ability to detect distant and putative novel IS elements while maintaining a moderate level of false positives. In this category it outperforms existing tools, especially when tested on large datasets of archaeal and bacterial genomes. **Conclusion:** We provide digIS, a software tool using a novel approach based on manually curated profile hidden Markov models, which is able to detect distant and putative novel IS elements. Although digIS can find known IS elements as well, we expect it to be used primarily by scientists interested in finding novel IS elements. The tool is available at `https://github.com/janka2012/digIS`.
Journal	BMC Bioinformatics, vol. 22, num. 258, 2021 Journal impact factor: 3.328, Q2
Citations	1 (WoS without self-citations)
Author's contribution	State-of-the-art study, algorithm design and implementation, tool testing, manuscript preparation.
Manuscript	Appendix A.10

Publication II

Title — Satellite DNA and Transposable Elements in Seabuckthorn (Hippophae rhamnoides), a Dioecious Plant with Small Y and Large X Chromosomes

Authors — PUTEROVÁ Janka, RAZUMOVA Olga, MARTÍNEK Tomáš, ALEXANDROV Oleg, DIVASHUK Mikhail, KUBÁT Zdeněk, HOBZA Roman, KARLOV Gennady, and KEJNOVSKÝ Eduard

Abstract — Seabuckthorn (Hippophae rhamnoides) is a dioecious shrub commonly used in the pharmaceutical, cosmetic, and environmental industry as a source of oil, minerals and vitamins. In this study,we analyzed the transposable elements and satellites in its genome.We carried out Illumina DNA sequencing and reconstructed the main repetitive DNA sequences. For data analysis, we developed a new bioinformatics approach for advanced satellite DNA analysis and showed that about 25% of the genome consists of satellite DNA and about 24% is formed of transposable elements, dominated by Ty3/Gypsy and Ty1/Copia LTR retrotransposons. FISH mapping revealed X chromosome-accumulated, Y chromosome-specific or both sex chromosomes-accumulated satellites butmost satellites were found on autosomes. Transposable elements were located mostly in the subtelomeres of all chromosomes. The 5S rDNA and 45S rDNA were localized on one autosomal locus each. Although we demonstrated the small size of the Y chromosome of the seabuckthorn and accumulated satellite DNA there, we were unable to estimate the age and extent of the Y chromosome degeneration. Analysis of dioecious relatives such as Shepherdia would shed more light on the evolution of these sex chromosomes.

Journal — Genome Biology and Evolution, vol. 9, num. 1, 2017
Journal impact factor: 3.940, Q1

Citations — 13 (WoS without self-citations)

Author's contribution — Consultation on the design and implementation of a new approach for satellite DNA analysis.

Manuscript — Appendix A.11

Publication III

Title Guanine quadruplexes are formed by specific regions of human transposable elements

Authors LEXA Matej, ŠTEFLOVÁ Pavlína, MARTÍNEK Tomáš, VORLÍČKOVÁ Michaela, VYSKOT Boris, and KEJNOVSKÝ Eduard

Abstract **Background:** Transposable elements form a significant proportion of eukaryotic genomes. Recently, Lexa et al. (Nucleic Acids Res 42:968-978, 2014) reported that plant long terminal repeat (LTR) retrotransposons often contain potential quadruplex sequences (PQSs) in their LTRs and experimentally confirmed their ability to adopt four-stranded DNA conformations.

Results: Here, we searched for PQSs in human retrotransposons and found that PQSs are specifically localized in the 3'-UTR of LINE-1 elements, in LTRs of HERV elements and are strongly accumulated in specific regions of SVA elements. Circular dichroism spectroscopy confirmed that most PQSs had adopted monomolecular or bimolecular guanine quadruplex structures. Evolutionarily young SVA elements contained more PQSs than older elements and their propensity to form quadruplex DNA was higher. Full-length L1 elements contained more PQSs than truncated elements; the highest proportion of PQSs was found inside transpositionally active L1 elements (PA2 and HS families).

Conclusion: Conservation of quadruplexes at specific positions of transposable elements implies their importance in their life cycle. The increasing quadruplex presence in evolutionarily young LINE-1 and SVA families makes these elements important contributors toward present genome-wide quadruplex distribution.

Journal BMC Genomics, vol. 15, num. 1032, 2014
Journal impact factor: 3.986, Q1

Citations 21 (WoS without self-citations)

Author's contribution Implementation of quadruplex search in the human genome.

Manuscript Appendix A.12

4.4 Conclusions

A growing number of studies are confirming the importance of repetitive sequences in the genomes of organisms and their essential role in many biological processes. To understand these relationships, it is essential to provide the scientific community with effective tools to search for and analyze them in assembled genomes and sequencing data.

Therefore, we have developed a new tool digIS, for searching Insertion Sequence elements in assembled genomes of prokaryotes. This tool can detect distant and putative novel IS elements. Although digIS can also find known IS elements, we expect it to be used primarily by scientists interested in finding novel IS elements and their experimental characterization.

The spectrum of tools for analyzing and quantifying repetitive sequences directly from sequencing data has been further extended by a new approach for processing satellite DNA sequences and their visualization. The developed technique was applied to the analysis of the seabuckthorn (Hippophae rhamnoides) genome and contributed to the interesting discovery of the Y-specific satellite showing that seabuckthorn has small Y and large X chromosomes since it was previously thought to be exactly the opposite.

Additionally, we also performed a study analyzing the relationships between transposable elements and specific secondary DNA structures, specifically quadruplexes. The study's results revealed several interesting insights into the occurrence of PQS within or in the vicinity of TEs, including their successful experimental evaluation in vitro.

4.4.1 Future work

The digIS tool for searching ISEs should be extended in two directions in the future:

- It can be expected that new families of IS elements will be discovered in the future, and the library of HMM profiles of catalytic domain models will need to be updated accordingly. It would therefore be beneficial to create a procedure to update this library automatically, for example, according to the ISfinder database, without the intervention of the original authors.
- During the experiments performed, we discovered several putative novel IS elements (see Additional file 9 of the original manuscript). It would be useful for digIS users if it also offered the possibility of additional analysis of these occurrences. For example, putative novel elements can further be clustered, aligned into a multiple sequence alignment, and verified for occurrence in other prokaryotic genomes.

In the case of the newly proposed approach for satellite DNA analysis and visualization, it would be useful to integrate it into the existing RepeatExplorer tool to extend its capabilities. Because of the importance of repeats in the study of genomes, several groups study this subject in parallel with us. As a result, a similar method called „TAndem REpeat ANalyzer“ (TAREAN) [141] was developed directly by the authors of RepeatExplorer and integrated into the new version of this tool.

4.5 Research impact

To make digIS easily accessible, its source code has been released as open-source on GitHub (`https://github.com/janka2012/digIS`). The installed tool, including all dependencies, is also available as a docker image at `https://hub.docker.com/r/janka2012/digis`. So far, we have registered 120 downloads of this tool and one citation. We attribute this relatively small number of citations to the only recent publication of the paper and also to the relatively small community of users dedicated to studying TEs in prokaryotic genomes.

In the case of the new approach for analyzing satellite DNA within the seabuckthorn genome, we recorded 13 citations. For the study of the relationships between TEs and quadruplexes, there was a total of 21 citations.

Bibliography

[1] UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Research*. vol. 47, no. D1. November 2018: pp. D506–D515. doi:10.1093/nar/gky1049.
Retrieved from: `https://doi.org/10.1093/nar/gky1049`

[2] Adámik, M.; Kejnovská, I.; Bažantová, P.; et al.: p53 binds human telomeric G-quadruplex in vitro. *Biochimie*. vol. 128-129. September 2016: pp. 83–91. doi:10.1016/j.biochi.2016.07.004.
Retrieved from: `https://doi.org/10.1016/j.biochi.2016.07.004`

[3] Adams, M. D.; Bishop, B.; Wright, M. S.: Quantitative assessment of insertion sequence impact on bacterial genome architecture. *Microbial Genomics*. vol. 2, no. 7. July 2016. doi:10.1099/mgen.0.000062.
Retrieved from: `https://doi.org/10.1099/mgen.0.000062`

[4] Agarwal, P.; States, D. J.: The Repeat Pattern Toolkit (RPT): Analyzing the Structure and Evolution of the C. elegans Genome. *Proceedings. International Conference on Intelligent Systems for Molecular Biology*. vol. 2. 1994: pp. 1–9.

[5] Agostini, F.; Vendruscolo, M.; Tartaglia, G. G.: Sequence-Based Prediction of Protein Solubility. *Journal of Molecular Biology*. vol. 421, no. 2-3. August 2012: pp. 237–241. doi:10.1016/j.jmb.2011.12.005.
Retrieved from: `https://doi.org/10.1016/j.jmb.2011.12.005`

[6] Alford, R. F.; Leaver-Fay, A.; Jeliazkov, J. R.; et al.: The Rosetta All-Atom Energy Function for Macromolecular Modeling and Design. *Journal of Chemical Theory and Computation*. vol. 13, no. 6. May 2017: pp. 3031–3048. doi:10.1021/acs.jctc.7b00125.
Retrieved from: `https://doi.org/10.1021/acs.jctc.7b00125`

[7] Altschul, S.: Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*. vol. 25, no. 17. September 1997: pp. 3389–3402. doi:10.1093/nar/25.17.3389.
Retrieved from: `https://doi.org/10.1093/nar/25.17.3389`

[8] Altschul, S. F.; Gish, W.; Miller, W.; et al.: Basic local alignment search tool. *Journal of Molecular Biology*. vol. 215, no. 3. October 1990: pp. 403–410. doi:10.1016/s0022-2836(05)80360-2.
Retrieved from: `https://doi.org/10.1016/s0022-2836(05)80360-2`

[9] Amin, N.; Liu, A.; Ramer, S.; et al.: Construction of stabilized proteins by combinatorial consensus mutagenesis. *Protein Engineering Design and Selection*. vol. 17, no. 11. November 2004: pp. 787–793. doi:10.1093/protein/gzh091.
Retrieved from: `https://doi.org/10.1093/protein/gzh091`

[10] Arakawa, T.; Timasheff, S. N.: [3]Theory of protein solubility. In *Methods in Enzymology*. Elsevier. 1985. pp. 49–77. doi:10.1016/0076-6879(85)14005-x. Retrieved from: `https://doi.org/10.1016/0076-6879(85)14005-x`

[11] Ashkenazy, H.; Penn, O.; Doron-Faigenboim, A.; et al.: FastML: a web server for probabilistic reconstruction of ancestral sequences. *Nucleic Acids Research*. vol. 40, no. W1. May 2012: pp. W580–W584. doi:10.1093/nar/gks498. Retrieved from: `https://doi.org/10.1093/nar/gks498`

[12] Babkova, P.; Sebestova, E.; Brezovsky, J.; et al.: Ancestral Haloalkane Dehalogenases Show Robustness and Unique Substrate Specificity. *ChemBioChem*. vol. 18, no. 14. June 2017: pp. 1448–1456. doi:10.1002/cbic.201700197. Retrieved from: `https://doi.org/10.1002/cbic.201700197`

[13] Bacolla, A.; Wells, R. D.: Non-B DNA conformations as determinants of mutagenesis and human disease. *Molecular Carcinogenesis*. vol. 48, no. 4. April 2009: pp. 273–285. doi:10.1002/mc.20507. Retrieved from: `https://doi.org/10.1002/mc.20507`

[14] Bao, W.; Kojima, K. K.; Kohany, O.: Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mobile DNA*. vol. 6, no. 1. June 2015. doi:10.1186/s13100-015-0041-9. Retrieved from: `https://doi.org/10.1186/s13100-015-0041-9`

[15] Bao, Z.; Eddy, S. R.: Automated De Novo Identification of Repeat Sequence Families in Sequenced Genomes. *Genome Research*. vol. 12, no. 8. July 2002: pp. 1269–1276. doi:10.1101/gr.88502. Retrieved from: `https://doi.org/10.1101/gr.88502`

[16] Barbara, M.: Mutable Loci in Maize. *Carnegie Institution of Washington Yearbook*. vol. 47. 1948: pp. 155–169.

[17] Barrett, T.; Clark, K.; Gevorgyan, R.; et al.: BioProject and BioSample databases at NCBI: facilitating capture and organization of metadata. *Nucleic Acids Research*. vol. 40, no. D1. December 2011: pp. D57–D63. doi:10.1093/nar/gkr1163. Retrieved from: `https://doi.org/10.1093/nar/gkr1163`

[18] Bednar, D.; Beerens, K.; Sebestova, E.; et al.: FireProt: Energy- and Evolution-Based Computational Design of Thermostable Multiple-Point Mutants. *PLOS Computational Biology*. vol. 11, no. 11. November 2015: page e1004556. doi:10.1371/journal.pcbi.1004556. Retrieved from: `https://doi.org/10.1371/journal.pcbi.1004556`

[19] Bedrat, A.; Lacroix, L.; Mergny, J.-L.: Re-evaluation of G-quadruplex propensity with G4Hunter. *Nucleic Acids Research*. vol. 44, no. 4. January 2016: pp. 1746–1759. doi:10.1093/nar/gkw006. Retrieved from: `https://doi.org/10.1093/nar/gkw006`

[20] Beerens, K.; Mazurenko, S.; Kunka, A.; et al.: Evolutionary Analysis As a Powerful Complement to Energy Calculations for Protein Stabilization. *ACS Catalysis*. vol. 8, no. 10. August 2018: pp. 9420–9428. doi:10.1021/acscatal.8b01677. Retrieved from: `https://doi.org/10.1021/acscatal.8b01677`

[21] Belotserkovskii, B. P.; Silva, E. D.; Tornaletti, S.; et al.: A Triplex-forming Sequence from the Human c-MYC Promoter Interferes with DNA Transcription. *Journal of Biological Chemistry*. vol. 282, no. 44. November 2007: pp. 32433–32441. doi:10.1074/jbc.m704618200.
Retrieved from: `https://doi.org/10.1074/jbc.m704618200`

[22] Bendl, J.; Stourac, J.; Sebestova, E.; et al.: HotSpot Wizard 2.0: automated design of site-specific mutations and smart libraries in protein engineering. *Nucleic Acids Research*. vol. 44, no. W1. May 2016: pp. W479–W487. doi:10.1093/nar/gkw416.
Retrieved from: `https://doi.org/10.1093/nar/gkw416`

[23] Benedix, A.; Becker, C. M.; de Groot, B. L.; et al.: Predicting free energy changes using structural ensembles. *Nature Methods*. vol. 6, no. 1. January 2009: pp. 3–4. doi:10.1038/nmeth0109-3.
Retrieved from: `https://doi.org/10.1038/nmeth0109-3`

[24] Benner, S. A.; Gerloff, D.: Patterns of divergence in homologous proteins as indicators of secondary and tertiary structure: A prediction of the structure of the catalytic domain of protein kinases. *Advances in Enzyme Regulation*. vol. 31. January 1991: pp. 121–181. doi:10.1016/0065-2571(91)90012-b.
Retrieved from: `https://doi.org/10.1016/0065-2571(91)90012-b`

[25] Bennetzen, J. L.; Wang, H.: The Contributions of Transposable Elements to the Structure, Function, and Evolution of Plant Genomes. *Annual Review of Plant Biology*. vol. 65, no. 1. April 2014: pp. 505–530. doi:10.1146/annurev-arplant-050213-035811.
Retrieved from: `https://doi.org/10.1146/annurev-arplant-050213-035811`

[26] Berman, H. M.: The Protein Data Bank. *Nucleic Acids Research*. vol. 28, no. 1. January 2000: pp. 235–242. doi:10.1093/nar/28.1.235.
Retrieved from: `https://doi.org/10.1093/nar/28.1.235`

[27] Berman, H. M.; Westbrook, J. D.; Gabanyi, M. J.; et al.: The protein structure initiative structural genomics knowledgebase. *Nucleic Acids Research*. vol. 37, no. Database. January 2009: pp. D365–D368. doi:10.1093/nar/gkn790.
Retrieved from: `https://doi.org/10.1093/nar/gkn790`

[28] Bhandari, B. K.; Gardner, P. P.; Lim, C. S.: Solubility-Weighted Index: fast and accurate prediction of protein solubility. *Bioinformatics*. vol. 36, no. 18. June 2020: pp. 4691–4698. doi:10.1093/bioinformatics/btaa578.
Retrieved from: `https://doi.org/10.1093/bioinformatics/btaa578`

[29] Biswas, A.; Gauthier, D. T.; Ranjan, D.; et al.: ISQuest: finding insertion sequences in prokaryotic sequence fragment data. *Bioinformatics*. vol. 31, no. 21. June 2015: pp. 3406–3412. doi:10.1093/bioinformatics/btv388.
Retrieved from: `https://doi.org/10.1093/bioinformatics/btv388`

[30] Bommarius, A. S.; Paye, M. F.: Stabilizing biocatalysts. *Chemical Society Reviews*. vol. 42, no. 15. 2013: page 6534. doi:10.1039/c3cs60137d.
Retrieved from: `https://doi.org/10.1039/c3cs60137d`

[31] Bornscheuer, U. T.; Huisman, G. W.; Kazlauskas, R. J.; et al.: Engineering the third wave of biocatalysis. *Nature*. vol. 485, no. 7397. May 2012: pp. 185–194. doi:10.1038/nature11117.
Retrieved from: `https://doi.org/10.1038/nature11117`

[32] Bowen, N. J.; Jordan, I. K.: Transposable Elements and the Evolution of Eukaryotic Complexity. *Current Issues in Molecular Biology*. vol. 4, no. 3. 2002: pp. 65–76. ISSN 1467-3045. doi:10.21775/cimb.004.065.
Retrieved from: `https://www.mdpi.com/1467-3045/4/3/7`

[33] Brannigan, J. A.; Wilkinson, A. J.: Protein engineering 20 years on. *Nature Reviews Molecular Cell Biology*. vol. 3, no. 12. December 2002: pp. 964–970. doi:10.1038/nrm975.
Retrieved from: `https://doi.org/10.1038/nrm975`

[34] Brenner, S.: The molecular evolution of genes and proteins: a tale of two serines. *Nature*. vol. 334, no. 6182. August 1988: pp. 528–530. doi:10.1038/334528a0.
Retrieved from: `https://doi.org/10.1038/334528a0`

[35] Brezovsky, J.; Chovancova, E.; Gora, A.; et al.: Software tools for identification, visualization and analysis of protein tunnels and channels. *Biotechnology Advances*. vol. 31, no. 1. January 2013: pp. 38–49. doi:10.1016/j.biotechadv.2012.02.002.
Retrieved from: `https://doi.org/10.1016/j.biotechadv.2012.02.002`

[36] Buchfink, B.; Reuter, K.; Drost, H.-G.: Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nature Methods*. vol. 18, no. 4. April 2021: pp. 366–368. doi:10.1038/s41592-021-01101-x.
Retrieved from: `https://doi.org/10.1038/s41592-021-01101-x`

[37] Buske, F. A.; Mattick, J. S.; Bailey, T. L.: Potential in vivo roles of nucleic acid triple-helices. *RNA Biology*. vol. 8, no. 3. May 2011: pp. 427–439. doi:10.4161/rna.8.3.14999.
Retrieved from: `https://doi.org/10.4161/rna.8.3.14999`

[38] Camacho, C.; Coulouris, G.; Avagyan, V.; et al.: BLAST+: architecture and applications. *BMC Bioinformatics*. vol. 10, no. 1. December 2009. doi:10.1186/1471-2105-10-421.
Retrieved from: `https://doi.org/10.1186/1471-2105-10-421`

[39] Capriotti, E.; Fariselli, P.; Casadio, R.: I-Mutant2.0: predicting stability changes upon mutation from the protein sequence or structure. *Nucleic Acids Research*. vol. 33, no. Web Server. July 2005: pp. W306–W310. doi:10.1093/nar/gki375.
Retrieved from: `https://doi.org/10.1093/nar/gki375`

[40] Cer, R. Z.; Bruce, K. H.; Mudunuri, U. S.; et al.: Non-B DB: a database of predicted non-B DNA-forming motifs in mammalian genomes. *Nucleic Acids Research*. vol. 39, no. Database. November 2010: pp. D383–D391. doi:10.1093/nar/gkq1170.
Retrieved from: `https://doi.org/10.1093/nar/gkq1170`

[41] Cerdobbel, A.; Winter, K. D.; Aerts, D.; et al.: Increasing the thermostability of sucrose phosphorylase by a combination of sequence- and structure-based mutagenesis. *Protein Engineering, Design and Selection*. vol. 24, no. 11. September

2011: pp. 829–834. doi:10.1093/protein/gzr042.
Retrieved from: `https://doi.org/10.1093/protein/gzr042`

[42] Chaloupková, R.; Sýkorová, J.; Prokop, Z.; et al.: Modification of Activity and Specificity of Haloalkane Dehalogenase from Sphingomonas paucimobilis UT26 by Engineering of Its Entrance Tunnel. *Journal of Biological Chemistry*. vol. 278, no. 52. December 2003: pp. 52622–52628. doi:10.1074/jbc.m306762200.
Retrieved from: `https://doi.org/10.1074/jbc.m306762200`

[43] Chambers, V. S.; Marsico, G.; Boutell, J. M.; et al.: High-throughput sequencing of DNA G-quadruplex structures in the human genome. *Nature Biotechnology*. vol. 33, no. 8. July 2015: pp. 877–881. doi:10.1038/nbt.3295.
Retrieved from: `https://doi.org/10.1038/nbt.3295`

[44] Chang, C. C. H.; Song, J.; Tey, B. T.; et al.: Bioinformatics approaches for improved recombinant protein production in Escherichia coli: protein solubility prediction. *Briefings in Bioinformatics*. vol. 15, no. 6. August 2013: pp. 953–962. doi:10.1093/bib/bbt057.
Retrieved from: `https://doi.org/10.1093/bib/bbt057`

[45] Chapman, J.; Ismail, A.; Dinu, C.: Industrial Applications of Enzymes: Recent Advances, Techniques, and Outlooks. *Catalysts*. vol. 8, no. 6. June 2018: page 238. doi:10.3390/catal8060238.
Retrieved from: `https://doi.org/10.3390/catal8060238`

[46] Charlesworth, D.: Plant sex chromosome evolution. *Journal of Experimental Botany*. vol. 64, no. 2. November 2012: pp. 405–420. doi:10.1093/jxb/ers322.
Retrieved from: `https://doi.org/10.1093/jxb/ers322`

[47] Chen, V. B.; Arendall, W. B.; Headd, J. J.; et al.: MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallographica Section D Biological Crystallography*. vol. 66, no. 1. December 2009: pp. 12–21. doi:10.1107/s0907444909042073.
Retrieved from: `https://doi.org/10.1107/s0907444909042073`

[48] Chénais, B.; Caruso, A.; Hiard, S.; et al.: The impact of transposable elements on eukaryotic genomes: From genome size increase to genetic adaptation to stressful environments. *Gene*. vol. 509, no. 1. November 2012: pp. 7–15. doi:10.1016/j.gene.2012.07.042.
Retrieved from: `https://doi.org/10.1016/j.gene.2012.07.042`

[49] Cheng, J.; Randall, A.; Baldi, P.: Prediction of protein stability changes for single-site mutations using support vector machines. *Proteins: Structure, Function, and Bioinformatics*. vol. 62, no. 4. December 2005: pp. 1125–1132. doi:10.1002/prot.20810.
Retrieved from: `https://doi.org/10.1002/prot.20810`

[50] Cho, J.; Paszkowski, J.: Regulation of rice root development by a retrotransposon acting as a microRNA sponge. *eLife*. vol. 6. August 2017. doi:10.7554/elife.30038.
Retrieved from: `https://doi.org/10.7554/elife.30038`

[51] Chu, C.; Nielsen, R.; Wu, Y.: REPdenovo: Inferring De Novo Repeat Motifs from Short Sequence Reads. *PLOS ONE*. vol. 11, no. 3. March 2016: page e0150719. doi:10.1371/journal.pone.0150719.
Retrieved from: `https://doi.org/10.1371/journal.pone.0150719`

[52] Cilia, E.; Pancsa, R.; Tompa, P.; et al.: The DynaMine webserver: predicting protein dynamics from sequence. *Nucleic Acids Research*. vol. 42, no. W1. April 2014: pp. W264–W270. doi:10.1093/nar/gku270.
Retrieved from: `https://doi.org/10.1093/nar/gku270`

[53] Cohen, S. N.; Chang, A. C. Y.; Boyer, H. W.; et al.: Construction of Biologically Functional Bacterial Plasmids In Vitro. *Proceedings of the National Academy of Sciences*. vol. 70, no. 11. November 1973: pp. 3240–3244. doi:10.1073/pnas.70.11.3240.
Retrieved from: `https://doi.org/10.1073/pnas.70.11.3240`

[54] Cooperman, B. S.; Baykov, A. A.; Lahti, R.: Evolutionary conservation of the active site of soluble inorganic pyrophosphatase. *Trends in Biochemical Sciences*. vol. 17, no. 7. July 1992: pp. 262–266. doi:10.1016/0968-0004(92)90406-y.
Retrieved from: `https://doi.org/10.1016/0968-0004(92)90406-y`

[55] Crooks, G. E.; Hon, G.; Chandonia, J.-M.; et al.: WebLogo: A Sequence Logo Generator: Figure 1. *Genome Research*. vol. 14, no. 6. June 2004: pp. 1188–1190. doi:10.1101/gr.849004.
Retrieved from: `https://doi.org/10.1101/gr.849004`

[56] Damborsky, J.: Meeting Report: Protein design and evolution for biocatalysis August 30 – September 1, 2006, Greifswald, Germany. *Biotechnology Journal*. vol. 2, no. 2. February 2007: pp. 176–179. doi:10.1002/biot.200600206.
Retrieved from: `https://doi.org/10.1002/biot.200600206`

[57] D'Antonio, L.; Bagga, P.: Computational methods for predicting intramolecular g-quadruplexes in nucleotide sequences. In *Proceedings. 2004 IEEE Computational Systems Bioinformatics Conference, 2004. CSB 2004.*. IEEE. doi:10.1109/csb.2004.1332508.
Retrieved from: `https://doi.org/10.1109/csb.2004.1332508`

[58] Davis, G. D.; Elisee, C.; Newham, D. M.; et al.: New fusion protein systems designed to give soluble expression inEscherichia coli. *Biotechnology and Bioengineering*. vol. 65, no. 4. November 1999: pp. 382–388. doi:10.1002/(sici)1097-0290(19991120)65:4<382::aid-bit2>3.0.co;2-i.
Retrieved from: `https://doi.org/10.1002/(sici)1097-0290(19991120)65:4<382::aid-bit2>3.0.co;2-i`

[59] Dehouck, Y.; Kwasigroch, J. M.; Gilis, D.; et al.: PoPMuSiC 2.1: a web server for the estimation of protein stability changes upon mutation and sequence optimality. *BMC Bioinformatics*. vol. 12, no. 1. May 2011. doi:10.1186/1471-2105-12-151.
Retrieved from: `https://doi.org/10.1186/1471-2105-12-151`

[60] Denard, C. A.; Ren, H.; Zhao, H.: Improving and repurposing biocatalysts via directed evolution. *Current Opinion in Chemical Biology*. vol. 25. April 2015: pp.

55–64. doi:10.1016/j.cbpa.2014.12.036.
Retrieved from: `https://doi.org/10.1016/j.cbpa.2014.12.036`

[61] Dewannieux, M.; Esnault, C.; Heidmann, T.: LINE-mediated retrotransposition of marked Alu sequences. *Nature Genetics*. vol. 35, no. 1. August 2003: pp. 41–48. doi:10.1038/ng1223.
Retrieved from: `https://doi.org/10.1038/ng1223`

[62] DEWANNIEUX, M.; HEIDMANN, T.: Role of poly(A) tail length in Alu retrotransposition. *Genomics*. vol. 86, no. 3. September 2005: pp. 378–381. doi:10.1016/j.ygeno.2005.05.009.
Retrieved from: `https://doi.org/10.1016/j.ygeno.2005.05.009`

[63] Dhapola, P.; Chowdhury, S.: QuadBase2: web server for multiplexed guanine quadruplex mining and visualization. *Nucleic Acids Research*. vol. 44, no. W1. May 2016: pp. W277–W283. doi:10.1093/nar/gkw425.
Retrieved from: `https://doi.org/10.1093/nar/gkw425`

[64] Diallo, A. B.; Makarenkov, V.; Blanchette, M.: Ancestors 1.0: a web server for ancestral sequence reconstruction. *Bioinformatics*. vol. 26, no. 1. October 2009: pp. 130–131. doi:10.1093/bioinformatics/btp600.
Retrieved from: `https://doi.org/10.1093/bioinformatics/btp600`

[65] Diaz, A. A.; Tomba, E.; Lennarson, R.; et al.: Prediction of protein solubility in Escherichia coli using logistic regression. *Biotechnology and Bioengineering*. vol. 105, no. 2. February 2010: pp. 374–383. doi:10.1002/bit.22537.
Retrieved from: `https://doi.org/10.1002/bit.22537`

[66] Dixon, B. P.; Lu, L.; Chu, A.; et al.: RecQ and RecG helicases have distinct roles in maintaining the stability of polypurine·polypyrimidine sequences. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*. vol. 643, no. 1-2. August 2008: pp. 20–28. doi:10.1016/j.mrfmmm.2008.05.005.
Retrieved from: `https://doi.org/10.1016/j.mrfmmm.2008.05.005`

[67] EDDY, S. R.: A NEW GENERATION OF HOMOLOGY SEARCH TOOLS BASED ON PROBABILISTIC INFERENCE. In *Genome Informatics 2009*. PUBLISHED BY IMPERIAL COLLEGE PRESS AND DISTRIBUTED BY WORLD SCIENTIFIC PUBLISHING CO.. October 2009. doi:10.1142/9781848165632_0019.
Retrieved from: `https://doi.org/10.1142/9781848165632_0019`

[68] Eddy, S. R.: Accelerated Profile HMM Searches. *PLoS Computational Biology*. vol. 7, no. 10. October 2011: page e1002195. doi:10.1371/journal.pcbi.1002195.
Retrieved from: `https://doi.org/10.1371/journal.pcbi.1002195`

[69] Edgar, R. C.: Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*. vol. 26, no. 19. August 2010: pp. 2460–2461. doi:10.1093/bioinformatics/btq461.
Retrieved from: `https://doi.org/10.1093/bioinformatics/btq461`

[70] Edgar, R. C.; Myers, E. W.: PILER: identification and classification of genomic repeats. *Bioinformatics*. vol. 21, no. Suppl 1. June 2005: pp. i152–i158.

doi:10.1093/bioinformatics/bti1003.
Retrieved from: `https://doi.org/10.1093/bioinformatics/bti1003`

[71] El-Deiry, W. S.; Kern, S. E.; Pietenpol, J. A.; et al.: Definition of a consensus binding site for p53. *Nature Genetics*. vol. 1, no. 1. April 1992: pp. 45–49. doi:10.1038/ng0492-45.
Retrieved from: `https://doi.org/10.1038/ng0492-45`

[72] Elena, T.; Capraru, G.; Rosu, C. M.; et al.: Morphometric pattern of somatic chromosomes in three Romanian seabuckthorn genotypes. *Caryologia*. vol. 64, no. 2. January 2011: pp. 189–196. doi:10.1080/00087114.2002.10589783.
Retrieved from: `https://doi.org/10.1080/00087114.2002.10589783`

[73] Federhen, S.: The NCBI Taxonomy database. *Nucleic Acids Research*. vol. 40, no. D1. December 2011: pp. D136–D143. doi:10.1093/nar/gkr1178.
Retrieved from: `https://doi.org/10.1093/nar/gkr1178`

[74] Feng, C.; Dai, M.; Liu, Y.; et al.: Sequence repetitiveness quantification and de novo repeat detection by weighted k-mer coverage. *Briefings in Bioinformatics*. vol. 22, no. 3. June 2020. doi:10.1093/bib/bbaa086.
Retrieved from: `https://doi.org/10.1093/bib/bbaa086`

[75] Fertin, G.; Jean, G.; Radulescu, A.; et al.: Hybrid de novo tandem repeat detection using short and long reads. *BMC Medical Genomics*. vol. 8, no. S3. September 2015. doi:10.1186/1755-8794-8-s3-s5.
Retrieved from: `https://doi.org/10.1186/1755-8794-8-s3-s5`

[76] Folkman, L.; Stantic, B.; Sattar, A.; et al.: EASE-MM: Sequence-Based Prediction of Mutation-Induced Stability Changes with Feature-Based Multiple Models. *Journal of Molecular Biology*. vol. 428, no. 6. March 2016: pp. 1394–1405. doi:10.1016/j.jmb.2016.01.012.
Retrieved from: `https://doi.org/10.1016/j.jmb.2016.01.012`

[77] Friedman, J. H.: Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*. vol. 29, no. 5. October 2001. doi:10.1214/aos/1013203451.
Retrieved from: `https://doi.org/10.1214/aos/1013203451`

[78] Gaddis, S. S.; Wu, Q.; Thames, H. D.; et al.: A Web-Based Search Engine for Triplex-forming Oligonucleotide Target Sequences. *Oligonucleotides*. vol. 16, no. 2. June 2006: pp. 196–201. doi:10.1089/oli.2006.16.196.
Retrieved from: `https://doi.org/10.1089/oli.2006.16.196`

[79] Gambin, T.; Startek, M.; Walczak, K.; et al.: TIRfinder: A Web Tool for Mining Class II Transposons Carrying Terminal Inverted Repeats. *Evolutionary Bioinformatics*. vol. 9. January 2013: page EBO.S10619. doi:10.4137/ebo.s10619.
Retrieved from: `https://doi.org/10.4137/ebo.s10619`

[80] Göbel, U.; Sander, C.; Schneider, R.; et al.: Correlated mutations and residue contacts in proteins. *Proteins: Structure, Function, and Genetics*. vol. 18, no. 4. April 1994: pp. 309–317. doi:10.1002/prot.340180402.
Retrieved from: `https://doi.org/10.1002/prot.340180402`

[81] Goldenzweig, A.; Goldsmith, M.; Hill, S. E.; et al.: Automated Structure- and Sequence-Based Design of Proteins for High Bacterial Expression and Stability. *Molecular Cell*. vol. 63, no. 2. July 2016: pp. 337–346. doi:10.1016/j.molcel.2016.06.012.
Retrieved from: `https://doi.org/10.1016/j.molcel.2016.06.012`

[82] Goubert, C.; Modolo, L.; Vieira, C.; et al.: De Novo Assembly and Annotation of the Asian Tiger Mosquito (Aedes albopictus) Repeatome with dnaPipeTE from Raw Genomic Reads and Comparative Analysis with the Yellow Fever Mosquito (Aedes aegypti). *Genome Biology and Evolution*. vol. 7, no. 4. March 2015: pp. 1192–1205. doi:10.1093/gbe/evv050.
Retrieved from: `https://doi.org/10.1093/gbe/evv050`

[83] Grabherr, M. G.; Haas, B. J.; Yassour, M.; et al.: Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology*. vol. 29, no. 7. May 2011: pp. 644–652. doi:10.1038/nbt.1883.
Retrieved from: `https://doi.org/10.1038/nbt.1883`

[84] Gromiha, M.: *Protein bioinformatics: from sequence to function*. Amsterdam Boston Paris: Elsevier. 2010. ISBN 9788131222973.

[85] Guo, R.; Li, Y.-R.; He, S.; et al.: RepLong: de novo repeat identification using long read sequencing data. *Bioinformatics*. vol. 34, no. 7. November 2017: pp. 1099–1107. doi:10.1093/bioinformatics/btx717.
Retrieved from: `https://doi.org/10.1093/bioinformatics/btx717`

[86] Haas, J.; Roth, S.; Arnold, K.; et al.: The Protein Model Portal—a comprehensive resource for protein structure and model information. *Database*. vol. 2013. January 2013. doi:10.1093/database/bat031.
Retrieved from: `https://doi.org/10.1093/database/bat031`

[87] Hawkey, J.; Hamidian, M.; Wick, R. R.; et al.: ISMapper: identifying transposase insertion sites in bacterial genomes from short read sequence data. *BMC Genomics*. vol. 16, no. 1. September 2015. doi:10.1186/s12864-015-1860-2.
Retrieved from: `https://doi.org/10.1186/s12864-015-1860-2`

[88] Hebditch, M.; Carballo-Amador, M. A.; Charonis, S.; et al.: Protein–Sol: a web tool for predicting protein solubility from sequence. *Bioinformatics*. vol. 33, no. 19. May 2017: pp. 3098–3100. doi:10.1093/bioinformatics/btx345.
Retrieved from: `https://doi.org/10.1093/bioinformatics/btx345`

[89] Hirose, S.; Noguchi, T.: ESPRESSO: A system for estimating protein expression and solubility in protein expression systems. *PROTEOMICS*. vol. 13, no. 9. April 2013: pp. 1444–1456. doi:10.1002/pmic.201200175.
Retrieved from: `https://doi.org/10.1002/pmic.201200175`

[90] Hooft, R. W. W.; Vriend, G.; Sander, C.; et al.: Errors in protein structures. *Nature*. vol. 381, no. 6580. May 1996: pp. 272–272. doi:10.1038/381272a0.
Retrieved from: `https://doi.org/10.1038/381272a0`

[91] Hoppe, C.; Schomburg, D.: Prediction of protein thermostability with a direction- and distance-dependent knowledge-based potential. *Protein Science*. vol. 14, no. 10.

October 2005: pp. 2682–2692. doi:10.1110/ps.04940705.
Retrieved from: https://doi.org/10.1110/ps.04940705

[92] Howell, N.: Evolutionary conservation of protein regions in the protonmotive cytochromeb and their possible roles in redox catalysis. *Journal of Molecular Evolution*. vol. 29, no. 2. August 1989: pp. 157–169. doi:10.1007/bf02100114.
Retrieved from: https://doi.org/10.1007/bf02100114

[93] Hoyne, P. R.; Edwards, L. M.; Viari, A.; et al.: Searching genomes for sequences with the potential to form intrastrand triple helices. *Journal of Molecular Biology*. vol. 302, no. 4. September 2000: pp. 797–809. doi:10.1006/jmbi.2000.4502.
Retrieved from: https://doi.org/10.1006/jmbi.2000.4502

[94] Huppert, J. L.: Prevalence of quadruplexes in the human genome. *Nucleic Acids Research*. vol. 33, no. 9. May 2005: pp. 2908–2916. doi:10.1093/nar/gki609.
Retrieved from: https://doi.org/10.1093/nar/gki609

[95] James, P. L.: Thermodynamic and kinetic stability of intermolecular triple helices containing different proportions of C+*GC and T*AT triplets. *Nucleic Acids Research*. vol. 31, no. 19. October 2003: pp. 5598–5606. doi:10.1093/nar/gkg782.
Retrieved from: https://doi.org/10.1093/nar/gkg782

[96] Jett, S. D.; Cherny, D. I.; Subramaniam, V.; et al.: Scanning force microscopy of the complexes of p53 core domain with supercoiled DNA 1 1Edited by M. Yaniv. *Journal of Molecular Biology*. vol. 299, no. 3. June 2000: pp. 585–592. doi:10.1006/jmbi.2000.3759.
Retrieved from: https://doi.org/10.1006/jmbi.2000.3759

[97] Jochens, H.; Aerts, D.; Bornscheuer, U. T.: Thermostabilization of an esterase by alignment-guided focussed directed evolution. *Protein Engineering, Design and Selection*. vol. 23, no. 12. October 2010: pp. 903–909. doi:10.1093/protein/gzq071.
Retrieved from: https://doi.org/10.1093/protein/gzq071

[98] KALYANARAMAN, A.; ALURU, S.: EFFICIENT ALGORITHMS AND SOFTWARE FOR DETECTION OF FULL-LENGTH LTR RETROTRANSPOSONS. *Journal of Bioinformatics and Computational Biology*. vol. 04, no. 02. April 2006: pp. 197–216. doi:10.1142/s021972000600203x.
Retrieved from: https://doi.org/10.1142/s021972000600203x

[99] Kang, S. M.; Wohlrab, F.; Wells, R. D.: Metal ions cause the isomerization of certain intramolecular triplexes. *Journal of Biological Chemistry*. vol. 267, no. 2. January 1992: pp. 1259–1264. doi:10.1016/s0021-9258(18)48423-2.
Retrieved from: https://doi.org/10.1016/s0021-9258(18)48423-2

[100] Karolchik, D.: The UCSC Table Browser data retrieval tool. *Nucleic Acids Research*. vol. 32, no. 90001. January 2004: pp. 493D–496. doi:10.1093/nar/gkh103.
Retrieved from: https://doi.org/10.1093/nar/gkh103

[101] Karolchik, D.; Barber, G. P.; Casper, J.; et al.: The UCSC Genome Browser database: 2014 update. *Nucleic Acids Research*. vol. 42, no. D1. November 2013: pp. D764–D770. doi:10.1093/nar/gkt1168.
Retrieved from: https://doi.org/10.1093/nar/gkt1168

[102] Karplus, M.; McCammon, J. A.: Molecular dynamics simulations of biomolecules. *Nature Structural Biology*. vol. 9, no. 9. September 2002: pp. 646–652. doi:10.1038/nsb0902-646.
Retrieved from: `https://doi.org/10.1038/nsb0902-646`

[103] Kellogg, E. H.; Leaver-Fay, A.; Baker, D.: Role of conformational sampling in computing mutation-induced changes in protein structure and stability. *Proteins: Structure, Function, and Bioinformatics*. vol. 79, no. 3. December 2010: pp. 830–838. doi:10.1002/prot.22921.
Retrieved from: `https://doi.org/10.1002/prot.22921`

[104] Khan, S.; Vihinen, M.: Performance of protein stability predictors. *Human Mutation*. vol. 31, no. 6. March 2010: pp. 675–684. doi:10.1002/humu.21242.
Retrieved from: `https://doi.org/10.1002/humu.21242`

[105] Khurana, S.; Rawi, R.; Kunji, K.; et al.: DeepSol: a deep learning framework for sequence-based protein solubility prediction. *Bioinformatics*. vol. 34, no. 15. March 2018: pp. 2605–2613. doi:10.1093/bioinformatics/bty166.
Retrieved from: `https://doi.org/10.1093/bioinformatics/bty166`

[106] Kichenaradja, P.; Siguier, P.; Pérochon, J.; et al.: ISbrowser: an extension of ISfinder for visualizing insertion sequences in prokaryotic genomes. *Nucleic Acids Research*. vol. 38, no. suppl_1. November 2009: pp. D62–D68. doi:10.1093/nar/gkp947.
Retrieved from: `https://doi.org/10.1093/nar/gkp947`

[107] Kikin, O.; D'Antonio, L.; Bagga, P. S.: QGRS Mapper: a web-based server for predicting G-quadruplexes in nucleotide sequences. *Nucleic Acids Research*. vol. 34, no. Web Server. July 2006: pp. W676–W682. doi:10.1093/nar/gkl253.
Retrieved from: `https://doi.org/10.1093/nar/gkl253`

[108] Klesmith, J. R.; Bacik, J.-P.; Wrenbeck, E. E.; et al.: Trade-offs between enzyme fitness and solubility illuminated by deep mutational scanning. *Proceedings of the National Academy of Sciences*. vol. 114, no. 9. February 2017: pp. 2265–2270. doi:10.1073/pnas.1614437114.
Retrieved from: `https://doi.org/10.1073/pnas.1614437114`

[109] Koch, P.; Platzer, M.; Downie, B. R.: RepARK—de novo creation of repeat libraries from whole-genome NGS reads. *Nucleic Acids Research*. vol. 42, no. 9. March 2014: pp. e80–e80. doi:10.1093/nar/gku210.
Retrieved from: `https://doi.org/10.1093/nar/gku210`

[110] de Koning, A. P. J.; Gu, W.; Castoe, T. A.; et al.: Repetitive Elements May Comprise Over Two-Thirds of the Human Genome. *PLoS Genetics*. vol. 7, no. 12. December 2011: page e1002384. doi:10.1371/journal.pgen.1002384.
Retrieved from: `https://doi.org/10.1371/journal.pgen.1002384`

[111] Krogh, A.; Larsson, B.; von Heijne, G.; et al.: Predicting transmembrane protein topology with a hidden markov model: application to complete genomes11Edited by F. Cohen. *Journal of Molecular Biology*. vol. 305, no. 3. January 2001: pp. 567–580. doi:10.1006/jmbi.2000.4315.
Retrieved from: `https://doi.org/10.1006/jmbi.2000.4315`

[112] Kuipers, R. K.; Joosten, H.-J.; van Berkel, W. J. H.; et al.: 3DM: Systematic analysis of heterogeneous superfamily data to discover protein functionalities. *Proteins: Structure, Function, and Bioinformatics*. 2010: pp. NA–NA. doi:10.1002/prot.22725.
Retrieved from: `https://doi.org/10.1002/prot.22725`

[113] Kumar, M. D. S.: ProTherm and ProNIT: thermodynamic databases for proteins and protein-nucleic acid interactions. *Nucleic Acids Research*. vol. 34, no. 90001. January 2006: pp. D204–D206. doi:10.1093/nar/gkj103.
Retrieved from: `https://doi.org/10.1093/nar/gkj103`

[114] Kurahashi, R.; ichi Tanaka, S.; Takano, K.: Activity-stability trade-off in random mutant proteins. *Journal of Bioscience and Bioengineering*. vol. 128, no. 4. October 2019: pp. 405–409. doi:10.1016/j.jbiosc.2019.03.017.
Retrieved from: `https://doi.org/10.1016/j.jbiosc.2019.03.017`

[115] Kurtz, S.; Schleiermacher, C.: REPuter: fast computation of maximal repeats in complete genomes. *Bioinformatics*. vol. 15, no. 5. May 1999: pp. 426–427. doi:10.1093/bioinformatics/15.5.426.
Retrieved from: `https://doi.org/10.1093/bioinformatics/15.5.426`

[116] Laimer, J.; Hofer, H.; Fritz, M.; et al.: MAESTRO - multi agent stability prediction upon point mutations. *BMC Bioinformatics*. vol. 16, no. 1. April 2015. doi:10.1186/s12859-015-0548-6.
Retrieved from: `https://doi.org/10.1186/s12859-015-0548-6`

[117] Laskowski, R. A.; MacArthur, M. W.; Moss, D. S.; et al.: PROCHECK: a program to check the stereochemical quality of protein structures. *Journal of Applied Crystallography*. vol. 26, no. 2. April 1993: pp. 283–291. doi:10.1107/s0021889892009944.
Retrieved from: `https://doi.org/10.1107/s0021889892009944`

[118] Lavecchia, A.; Giovanni, C.: Virtual Screening Strategies in Drug Discovery: A Critical Review. *Current Medicinal Chemistry*. vol. 20, no. 23. June 2013: pp. 2839–2860. doi:10.2174/09298673113209990001.
Retrieved from: `https://doi.org/10.2174/09298673113209990001`

[119] Lehmann, M.; Loch, C.; Middendorf, A.; et al.: The consensus concept for thermostability engineering of proteins: further proof of concept. *Protein Engineering, Design and Selection*. vol. 15, no. 5. May 2002: pp. 403–411. doi:10.1093/protein/15.5.403.
Retrieved from: `https://doi.org/10.1093/protein/15.5.403`

[120] Lexa, M.; Steflova, P.; Martinek, T.; et al.: Guanine quadruplexes are formed by specific regions of human transposable elements. *BMC Genomics*. vol. 15, no. 1. November 2014. doi:10.1186/1471-2164-15-1032.
Retrieved from: `https://doi.org/10.1186/1471-2164-15-1032`

[121] Li, H.: Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. 2013. doi:10.48550/ARXIV.1303.3997.
Retrieved from: `https://arxiv.org/abs/1303.3997`

[122] Li, R.; Ye, J.; Li, S.; et al.: ReAS: Recovery of Ancestral Sequences for Transposable Elements from the Unassembled Reads of a Whole Genome Shotgun. *PLoS Computational Biology*. vol. 1, no. 4. September 2005: page e43. doi:10.1371/journal.pcbi.0010043.
Retrieved from: `https://doi.org/10.1371/journal.pcbi.0010043`

[123] Li, S.-F.; Su, T.; Cheng, G.-Q.; et al.: Chromosome Evolution in Connection with Repetitive Sequences and Epigenetics in Plants. *Genes*. vol. 8, no. 10. October 2017: page 290. doi:10.3390/genes8100290.
Retrieved from: `https://doi.org/10.3390/genes8100290`

[124] Li, W.; O'Neill, K. R.; Haft, D. H.; et al.: RefSeq: expanding the Prokaryotic Genome Annotation Pipeline reach with protein family model curation. *Nucleic Acids Research*. vol. 49, no. D1. December 2020: pp. D1020–D1028. doi:10.1093/nar/gkaa1105.
Retrieved from: `https://doi.org/10.1093/nar/gkaa1105`

[125] Li, Y.; Fang, J.: PROTS-RF: A Robust Model for Predicting Mutation-Induced Protein Stability Changes. *PLoS ONE*. vol. 7, no. 10. October 2012: page e47247. doi:10.1371/journal.pone.0047247.
Retrieved from: `https://doi.org/10.1371/journal.pone.0047247`

[126] Liao, X.; Zhang, X.; Wu, F.-X.; et al.: de novo repeat detection based on the third generation sequencing reads. In *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE. November 2019. doi:10.1109/bibm47256.2019.8982959.
Retrieved from: `https://doi.org/10.1109/bibm47256.2019.8982959`

[127] Lu, G.: Vector NTI, a balanced all-in-one sequence analysis suite. *Briefings in Bioinformatics*. vol. 5, no. 4. January 2004: pp. 378–388. doi:10.1093/bib/5.4.378.
Retrieved from: `https://doi.org/10.1093/bib/5.4.378`

[128] Macas, J.; Meszaros, T.; Nouzova, M.: PlantSat: a specialized database for plant satellite repeats. *Bioinformatics*. vol. 18, no. 1. January 2002: pp. 28–35. doi:10.1093/bioinformatics/18.1.28.
Retrieved from: `https://doi.org/10.1093/bioinformatics/18.1.28`

[129] Magnan, C. N.; Randall, A.; Baldi, P.: SOLpro: accurate sequence-based prediction of protein solubility. *Bioinformatics*. vol. 25, no. 17. June 2009: pp. 2200–2207. doi:10.1093/bioinformatics/btp386.
Retrieved from: `https://doi.org/10.1093/bioinformatics/btp386`

[130] Martin, A.; Troadec, C.; Boualem, A.; et al.: A transposon-induced epigenetic change leads to sex determination in melon. *Nature*. vol. 461, no. 7267. October 2009: pp. 1135–1138. doi:10.1038/nature08498.
Retrieved from: `https://doi.org/10.1038/nature08498`

[131] McCarthy, E. M.; McDonald, J. F.: LTR_STRUC: a novel search and identification program for LTR retrotransposons. *Bioinformatics*. vol. 19, no. 3. February 2003: pp. 362–367. doi:10.1093/bioinformatics/btf878.
Retrieved from: `https://doi.org/10.1093/bioinformatics/btf878`

[132] Mergny, J. L.; Sun, J. S.; Rougee, M.; et al.: Sequence specificity in triple helix formation: experimental and theoretical studies of the effect of mismatches on triplex stability. *Biochemistry*. vol. 30, no. 40. October 1991: pp. 9791–9798. doi:10.1021/bi00104a031.
Retrieved from: `https://doi.org/10.1021/bi00104a031`

[133] Michael, T. P.: Plant genome size variation: bloating and purging DNA. *Briefings in Functional Genomics*. vol. 13, no. 4. March 2014: pp. 308–317. doi:10.1093/bfgp/elu005.
Retrieved from: `https://doi.org/10.1093/bfgp/elu005`

[134] Miga, K. H.: Completing the human genome: the progress and challenge of satellite DNA assembly. *Chromosome Research*. vol. 23, no. 3. September 2015: pp. 421–426. doi:10.1007/s10577-015-9488-2.
Retrieved from: `https://doi.org/10.1007/s10577-015-9488-2`

[135] Mistry, J.; Finn, R. D.; Eddy, S. R.; et al.: Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. *Nucleic Acids Research*. vol. 41, no. 12. April 2013: pp. e121–e121. doi:10.1093/nar/gkt263.
Retrieved from: `https://doi.org/10.1093/nar/gkt263`

[136] Mitchell, A. L.; Almeida, A.; Beracochea, M.; et al.: MGnify: the microbiome analysis resource in 2020. *Nucleic Acids Research*. November 2019. doi:10.1093/nar/gkz1035.
Retrieved from: `https://doi.org/10.1093/nar/gkz1035`

[137] Morley, K. L.; Kazlauskas, R. J.: Improving enzyme properties: when are closer mutations better? *Trends in Biotechnology*. vol. 23, no. 5. May 2005: pp. 231–237. doi:10.1016/j.tibtech.2005.03.005.
Retrieved from: `https://doi.org/10.1016/j.tibtech.2005.03.005`

[138] Mukundan, V. T.; Phan, A. T.: Bulges in G-Quadruplexes: Broadening the Definition of G-Quadruplex-Forming Sequences. *Journal of the American Chemical Society*. vol. 135, no. 13. March 2013: pp. 5017–5028. doi:10.1021/ja310251r.
Retrieved from: `https://doi.org/10.1021/ja310251r`

[139] Neher, E.: How frequent are correlated changes in families of protein sequences? *Proceedings of the National Academy of Sciences*. vol. 91, no. 1. January 1994: pp. 98–102. doi:10.1073/pnas.91.1.98.
Retrieved from: `https://doi.org/10.1073/pnas.91.1.98`

[140] Novak, P.; Neumann, P.; Pech, J.; et al.: RepeatExplorer: a Galaxy-based web server for genome-wide characterization of eukaryotic repetitive elements from next-generation sequence reads. *Bioinformatics*. vol. 29, no. 6. February 2013: pp. 792–793. doi:10.1093/bioinformatics/btt054.
Retrieved from: `https://doi.org/10.1093/bioinformatics/btt054`

[141] Novák, P.; Robledillo, L. Á.; Koblížková, A.; et al.: TAREAN: a computational tool for identification and characterization of satellite DNA from unassembled short reads. *Nucleic Acids Research*. vol. 45, no. 12. April 2017: pp. e111–e111. doi:10.1093/nar/gkx257.
Retrieved from: `https://doi.org/10.1093/nar/gkx257`

[142] Novotna, B.; Mares, J.: *Vyvojova biologie pro mediky*. Praha: Karolinum. first edition. 2005. ISBN 9788024610238. oCLC: 85163514.

[143] Oliveira, P. H.; Prazeres, D.; Monteiro, G.: *DNA instability in bacterial genomes: causes and consequences*. 01 2014. ISBN 978-1-908230-29-4. pp. 261–284.

[144] Parthiban, V.; Gromiha, M. M.; Schomburg, D.: CUPSAT: prediction of protein stability upon point mutations. *Nucleic Acids Research*. vol. 34, no. Web Server. July 2006: pp. W239–W242. doi:10.1093/nar/gkl190.
Retrieved from: `https://doi.org/10.1093/nar/gkl190`

[145] Pérez, A.; Marchán, I.; Svozil, D.; et al.: Refinement of the AMBER Force Field for Nucleic Acids: Improving the Description of alpha/gamma Conformers. *Biophysical Journal*. vol. 92, no. 11. June 2007: pp. 3817–3829. doi:10.1529/biophysj.106.097782.
Retrieved from: `https://doi.org/10.1529/biophysj.106.097782`

[146] Pey, A. L.; Rodriguez-Larrea, D.; Bomke, S.; et al.: Engineering proteins with tunable thermodynamic and kinetic stabilities. *Proteins: Structure, Function, and Bioinformatics*. vol. 71, no. 1. April 2008: pp. 165–174. doi:10.1002/prot.21670.
Retrieved from: `https://doi.org/10.1002/prot.21670`

[147] Pezer, Ž.; Brajković, J.; Feliciello, I.; et al.: Satellite DNA-Mediated Effects on Genome Regulation. In *Genome Dynamics*. S. Karger AG. 2012. pp. 153–169. doi:10.1159/000337116.
Retrieved from: `https://doi.org/10.1159/000337116`

[148] Piovesan, D.; Walsh, I.; Minervini, G.; et al.: FELLS: fast estimator of latent local structure. *Bioinformatics*. vol. 33, no. 12. February 2017: pp. 1889–1891. doi:10.1093/bioinformatics/btx085.
Retrieved from: `https://doi.org/10.1093/bioinformatics/btx085`

[149] Plohl, M.; Meštrović, N.; Mravinac, B.: Satellite DNA Evolution. In *Genome Dynamics*. S. Karger AG. 2012. pp. 126–152. doi:10.1159/000337122.
Retrieved from: `https://doi.org/10.1159/000337122`

[150] Plum, G. E.; Pilch, D. S.; Singleton, S. F.; et al.: Nucleic Acid Hybridization: Triplex Stability and Energetics. *Annual Review of Biophysics and Biomolecular Structure*. vol. 24, no. 1. June 1995: pp. 319–350. doi:10.1146/annurev.bb.24.060195.001535.
Retrieved from: `https://doi.org/10.1146/annurev.bb.24.060195.001535`

[151] Polizzi, K. M.; Bommarius, A. S.; Broering, J. M.; et al.: Stability of biocatalysts. *Current Opinion in Chemical Biology*. vol. 11, no. 2. April 2007: pp. 220–225. doi:10.1016/j.cbpa.2007.01.685.
Retrieved from: `https://doi.org/10.1016/j.cbpa.2007.01.685`

[152] Potapov, V.; Cohen, M.; Schreiber, G.: Assessing computational methods for predicting protein stability upon mutation: good on average but not in the details. *Protein Engineering Design and Selection*. vol. 22, no. 9. June 2009: pp. 553–560. doi:10.1093/protein/gzp030.
Retrieved from: `https://doi.org/10.1093/protein/gzp030`

[153] Price, A. L.; Jones, N. C.; Pevzner, P. A.: De novo identification of repeat families in large genomes. *Bioinformatics*. vol. 21, no. Suppl 1. June 2005: pp. i351–i358. doi:10.1093/bioinformatics/bti1018.
Retrieved from: `https://doi.org/10.1093/bioinformatics/bti1018`

[154] Pucci, F.; Bernaerts, K. V.; Kwasigroch, J. M.; et al.: Quantification of biases in predictions of protein stability changes upon mutations. *Bioinformatics*. vol. 34, no. 21. April 2018: pp. 3659–3665. doi:10.1093/bioinformatics/bty348.
Retrieved from: `https://doi.org/10.1093/bioinformatics/bty348`

[155] Pucci, F.; Bourgeas, R.; Rooman, M.: Predicting protein thermal stability changes upon point mutations using statistical potentials: Introducing HoTMuSiC. *Scientific Reports*. vol. 6, no. 1. March 2016. doi:10.1038/srep23257.
Retrieved from: `https://doi.org/10.1038/srep23257`

[156] Puterova, J.; Razumova, O.; Martinek, T.; et al.: Satellite DNA and Transposable Elements in Seabuckthorn (Hippophae rhamnoides), a Dioecious Plant with Small Y and Large X Chromosomes. *Genome Biology and Evolution*. January 2017: page evw303. doi:10.1093/gbe/evw303.
Retrieved from: `https://doi.org/10.1093/gbe/evw303`

[157] Quevillon, E.; Silventoinen, V.; Pillai, S.; et al.: InterProScan: protein domains identifier. *Nucleic Acids Research*. vol. 33, no. Web Server. July 2005: pp. W116–W120. doi:10.1093/nar/gki442.
Retrieved from: `https://doi.org/10.1093/nar/gki442`

[158] Raghavan, S. C.; Chastain, P.; Lee, J. S.; et al.: Evidence for a Triplex DNA Conformation at the bcl-2 Major Breakpoint Region of the t(14;18) Translocation*. *Journal of Biological Chemistry*. vol. 280, no. 24. 2005: pp. 22749–22760. ISSN 0021-9258. doi:https://doi.org/10.1074/jbc.M502952200.
Retrieved from:
`https://www.sciencedirect.com/science/article/pii/S0021925820614429`

[159] Raimondi, D.; Orlando, G.; Fariselli, P.; et al.: Insight into the protein solubility driving forces with neural attention. *PLOS Computational Biology*. vol. 16, no. 4. April 2020: page e1007722. doi:10.1371/journal.pcbi.1007722.
Retrieved from: `https://doi.org/10.1371/journal.pcbi.1007722`

[160] Rathinavelan, T.; Yathindra, N.: Base triplet nonisomorphism strongly influences DNA triplex conformation: Effect of nonisomorphic G*GC and A*AT triplets and bending of DNA triplexes. *Biopolymers*. vol. 82, no. 5. August 2006: pp. 443–461. doi:10.1002/bip.20484.
Retrieved from: `https://doi.org/10.1002/bip.20484`

[161] Reetz, M. T.; Bocola, M.; Carballeira, J. D.; et al.: Expanding the Range of Substrate Acceptance of Enzymes: Combinatorial Active-Site Saturation Test. *Angewandte Chemie International Edition*. vol. 44, no. 27. July 2005: pp. 4192–4196. doi:10.1002/anie.200500767.
Retrieved from: `https://doi.org/10.1002/anie.200500767`

[162] Reetz, M. T.; Carballeira, J. D.; Vogel, A.: Iterative Saturation Mutagenesis on the Basis of B Factors as a Strategy for Increasing Protein Thermostability. *Angewandte*

Chemie International Edition. vol. 45, no. 46. November 2006: pp. 7745–7751. doi:10.1002/anie.200602795.
Retrieved from: `https://doi.org/10.1002/anie.200602795`

[163] Reetz, M. T.; Torre, C.; Eipper, A.; et al.: Enhancing the Enantioselectivity of an Epoxide Hydrolase by Directed Evolution. *Organic Letters*. vol. 6, no. 2. December 2003: pp. 177–180. doi:10.1021/ol035898m.
Retrieved from: `https://doi.org/10.1021/ol035898m`

[164] Rembeza, E.; Engqvist, M. K.: Experimental investigation of enzyme functional annotations reveals extensive annotation error. December 2020. doi:10.1101/2020.12.18.423474.
Retrieved from: `https://doi.org/10.1101/2020.12.18.423474`

[165] Rhodes, D.; Lipps, H. J.: G-quadruplexes and their regulatory roles in biology. *Nucleic Acids Research*. vol. 43, no. 18. September 2015: pp. 8627–8637. doi:10.1093/nar/gkv862.
Retrieved from: `https://doi.org/10.1093/nar/gkv862`

[166] Riadi, G.; Medina-Moenne, C.; Holmes, D. S.: TnpPred: A Web Service for the Robust Prediction of Prokaryotic Transposases. *Comparative and Functional Genomics*. vol. 2012. 2012: pp. 1–5. doi:10.1155/2012/678761.
Retrieved from: `https://doi.org/10.1155/2012/678761`

[167] Rice, P.; Longden, I.; Bleasby, A.: EMBOSS: The European Molecular Biology Open Software Suite. *Trends in Genetics*. vol. 16, no. 6. June 2000: pp. 276–277. doi:10.1016/s0168-9525(00)02024-2.
Retrieved from: `https://doi.org/10.1016/s0168-9525(00)02024-2`

[168] Rippe, K.; Fritsch, V.; Westhof, E.; et al.: Alternating d(G-A) sequences form a parallel-stranded DNA homoduplex. *The EMBO Journal*. vol. 11, no. 10. October 1992: pp. 3777–3786. doi:10.1002/j.1460-2075.1992.tb05463.x.
Retrieved from: `https://doi.org/10.1002/j.1460-2075.1992.tb05463.x`

[169] Roberts, R. W.; Crothers, D. M.: Specificity and stringency in DNA triplex formation. *Proceedings of the National Academy of Sciences*. vol. 88, no. 21. November 1991: pp. 9397–9401. doi:10.1073/pnas.88.21.9397.
Retrieved from: `https://doi.org/10.1073/pnas.88.21.9397`

[170] Robinson, D. G.; Lee, M.-C.; Marx, C. J.: OASIS: an automated program for global investigation of bacterial and archaeal insertion sequences. *Nucleic Acids Research*. vol. 40, no. 22. August 2012: pp. e174–e174. doi:10.1093/nar/gks778.
Retrieved from: `https://doi.org/10.1093/nar/gks778`

[171] Ronquist, F.; Teslenko, M.; van der Mark, P.; et al.: MrBayes 3.2: Efficient Bayesian Phylogenetic Inference and Model Choice Across a Large Model Space. *Systematic Biology*. vol. 61, no. 3. February 2012: pp. 539–542. doi:10.1093/sysbio/sys029.
Retrieved from: `https://doi.org/10.1093/sysbio/sys029`

[172] Roy-Engel, A. M.: A tale of an A-tail. *Mobile Genetic Elements*. vol. 2, no. 6. November 2012: pp. 282–286. doi:10.4161/mge.23204.
Retrieved from: `https://doi.org/10.4161/mge.23204`

[173] Ruiz-Ruano, F. J.; López-León, M. D.; Cabrero, J.; et al.: High-throughput analysis of the satellitome illuminates satellite DNA evolution. *Scientific Reports*. vol. 6, no. 1. July 2016. doi:10.1038/srep28333.
Retrieved from: `https://doi.org/10.1038/srep28333`

[174] Salomon-Ferrer, R.; Case, D. A.; Walker, R. C.: An overview of the Amber biomolecular simulation package. *Wiley Interdisciplinary Reviews: Computational Molecular Science*. vol. 3, no. 2. September 2012: pp. 198–210. doi:10.1002/wcms.1121.
Retrieved from: `https://doi.org/10.1002/wcms.1121`

[175] SantaLucia, J.: A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics. *Proceedings of the National Academy of Sciences*. vol. 95, no. 4. February 1998: pp. 1460–1465. doi:10.1073/pnas.95.4.1460.
Retrieved from: `https://doi.org/10.1073/pnas.95.4.1460`

[176] Sayers, E. W.; Agarwala, R.; Bolton, E. E.; et al.: Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research*. vol. 47, no. D1. November 2018: pp. D23–D28. doi:10.1093/nar/gky1069.
Retrieved from: `https://doi.org/10.1093/nar/gky1069`

[177] Scaria, V.; Hariharan, M.; Arora, A.; et al.: Quadfinder: server for identification and analysis of quadruplex-forming motifs in nucleotide sequences. *Nucleic Acids Research*. vol. 34, no. Web Server. July 2006: pp. W683–W685. doi:10.1093/nar/gkl299.
Retrieved from: `https://doi.org/10.1093/nar/gkl299`

[178] Schnable, P. S.; Ware, D.; Fulton, R. S.; et al.: The B73 Maize Genome: Complexity, Diversity, and Dynamics. *Science*. vol. 326, no. 5956. November 2009: pp. 1112–1115. doi:10.1126/science.1178534.
Retrieved from: `https://doi.org/10.1126/science.1178534`

[179] Schroth, G. P.; Ho, P. S.: Occurrence of potential cruciform and H-DNA forming sequences in genomic DNA. *Nucleic Acids Research*. vol. 23, no. 11. 1995: pp. 1977–1983. doi:10.1093/nar/23.11.1977.
Retrieved from: `https://doi.org/10.1093/nar/23.11.1977`

[180] Schymkowitz, J.; Borg, J.; Stricher, F.; et al.: The FoldX web server: an online force field. *Nucleic Acids Research*. vol. 33, no. Web Server. July 2005: pp. W382–W388. doi:10.1093/nar/gki387.
Retrieved from: `https://doi.org/10.1093/nar/gki387`

[181] Scrucca, L.: GA: A Package for Genetic Algorithms in R. *Journal of Statistical Software*. vol. 53, no. 4. 2013. doi:10.18637/jss.v053.i04.
Retrieved from: `https://doi.org/10.18637/jss.v053.i04`

[182] Sebestova, E.; Bendl, J.; Brezovsky, J.; et al.: Computational Tools for Designing Smart Libraries. In *Methods in Molecular Biology*. Springer New York. 2014. pp. 291–314. doi:10.1007/978-1-4939-1053-3_20.
Retrieved from: `https://doi.org/10.1007/978-1-4939-1053-3_20`

[183] Seidman, M. M.; Glazer, P. M.: The potential for gene repair via triple helix formation. *Journal of Clinical Investigation*. vol. 112, no. 4. August 2003: pp. 487–494. doi:10.1172/jci19552.
Retrieved from: `https://doi.org/10.1172/jci19552`

[184] Siddiqui, K. S.: Defying the activity–stability trade-off in enzymes: taking advantage of entropy to enhance activity and thermostability. *Critical Reviews in Biotechnology*. vol. 37, no. 3. March 2016: pp. 309–322. doi:10.3109/07388551.2016.1144045.
Retrieved from: `https://doi.org/10.3109/07388551.2016.1144045`

[185] Sievers, F.; Wilm, A.; Dineen, D.; et al.: Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular Systems Biology*. vol. 7, no. 1. January 2011: page 539. doi:10.1038/msb.2011.75.
Retrieved from: `https://doi.org/10.1038/msb.2011.75`

[186] Siguier, P.: ISfinder: the reference centre for bacterial insertion sequences. *Nucleic Acids Research*. vol. 34, no. 90001. January 2006: pp. D32–D36. doi:10.1093/nar/gkj014.
Retrieved from: `https://doi.org/10.1093/nar/gkj014`

[187] Siguier, P.; Gourbeyre, E.; Chandler, M.: Bacterial insertion sequences: their genomic impact and diversity. *FEMS Microbiology Reviews*. vol. 38, no. 5. September 2014: pp. 865–891. doi:10.1111/1574-6976.12067.
Retrieved from: `https://doi.org/10.1111/1574-6976.12067`

[188] Simpson, J. T.; Wong, K.; Jackman, S. D.; et al.: ABySS: A parallel assembler for short read sequence data. *Genome Research*. vol. 19, no. 6. February 2009: pp. 1117–1123. doi:10.1101/gr.089532.108.
Retrieved from: `https://doi.org/10.1101/gr.089532.108`

[189] Smialowski, P.; Doose, G.; Torkler, P.; et al.: PROSO II - a new method for protein solubility prediction. *FEBS Journal*. vol. 279, no. 12. May 2012: pp. 2192–2200. doi:10.1111/j.1742-4658.2012.08603.x.
Retrieved from: `https://doi.org/10.1111/j.1742-4658.2012.08603.x`

[190] Smith, T.; Waterman, M.: Identification of common molecular subsequences. *Journal of Molecular Biology*. vol. 147, no. 1. March 1981: pp. 195–197. doi:10.1016/0022-2836(81)90087-5.
Retrieved from: `https://doi.org/10.1016/0022-2836(81)90087-5`

[191] Smyth, M. S.: x Ray crystallography. *Molecular Pathology*. vol. 53, no. 1. February 2000: pp. 8–14. doi:10.1136/mp.53.1.8.
Retrieved from: `https://doi.org/10.1136/mp.53.1.8`

[192] Sokal, R. R.; Michener, C. D.: A statistical method for evaluating systematic relationships. *University of Kansas science bulletin*. vol. 38. 1958: pp. 1409–1438.

[193] Sormanni, P.; Aprile, F. A.; Vendruscolo, M.: The CamSol Method of Rational Design of Protein Mutants with Enhanced Solubility. *Journal of Molecular Biology*. vol. 427, no. 2. January 2015: pp. 478–490. doi:10.1016/j.jmb.2014.09.026.
Retrieved from: `https://doi.org/10.1016/j.jmb.2014.09.026`

[194] Stamatakis, A.: RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics.* vol. 30, no. 9. January 2014: pp. 1312–1313. doi:10.1093/bioinformatics/btu033.
Retrieved from: `https://doi.org/10.1093/bioinformatics/btu033`

[195] Staton, S. E.; Burke, J. M.: Transposome: a toolkit for annotation of transposable element families from unassembled sequence reads. *Bioinformatics.* vol. 31, no. 11. February 2015: pp. 1827–1829. doi:10.1093/bioinformatics/btv059.
Retrieved from: `https://doi.org/10.1093/bioinformatics/btv059`

[196] Steinbiss, S.; Willhoeft, U.; Gremme, G.; et al.: Fine-grained annotation and classification of de novo predicted LTR retrotransposons. *Nucleic Acids Research.* vol. 37, no. 21. September 2009: pp. 7002–7013. doi:10.1093/nar/gkp759.
Retrieved from: `https://doi.org/10.1093/nar/gkp759`

[197] Steinegger, M.; Söding, J.: MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature Biotechnology.* vol. 35, no. 11. October 2017: pp. 1026–1028. doi:10.1038/nbt.3988.
Retrieved from: `https://doi.org/10.1038/nbt.3988`

[198] Štros, M.; Muselíková-Polanská, E.; Pospíšilová, Š.; et al.: High-Affinity Binding of Tumor-Suppressor Protein p53 and HMGB1 to Hemicatenated DNA Loops. *Biochemistry.* vol. 43, no. 22. May 2004: pp. 7215–7225. doi:10.1021/bi049928k.
Retrieved from: `https://doi.org/10.1021/bi049928k`

[199] Sullivan, B. J.; Nguyen, T.; Durani, V.; et al.: Stabilizing Proteins from Sequence Statistics: The Interplay of Conservation and Correlation in Triosephosphate Isomerase Stability. *Journal of Molecular Biology.* vol. 420, no. 4-5. July 2012: pp. 384–399. doi:10.1016/j.jmb.2012.04.025.
Retrieved from: `https://doi.org/10.1016/j.jmb.2012.04.025`

[200] Tan, Z.-J.; Chen, S.-J.: Nucleic Acid Helix Stability: Effects of Salt Concentration, Cation Valence and Size, and Chain Length. *Biophysical Journal.* vol. 90, no. 4. February 2006: pp. 1175–1190. doi:10.1529/biophysj.105.070904.
Retrieved from: `https://doi.org/10.1529/biophysj.105.070904`

[201] Taylor, W. R.; Hatrick, K.: Compensating changes in protein multiple sequence alignments. *„Protein Engineering, Design and Selection“.* vol. 7, no. 3. 1994: pp. 341–348. doi:10.1093/protein/7.3.341.
Retrieved from: `https://doi.org/10.1093/protein/7.3.341`

[202] Teng, S.; Srivastava, A. K.; Wang, L.: Sequence feature-based prediction of protein stability changes upon amino acid substitutions. *BMC Genomics.* vol. 11, no. Suppl 2. 2010: page S5. doi:10.1186/1471-2164-11-s2-s5.
Retrieved from: `https://doi.org/10.1186/1471-2164-11-s2-s5`

[203] Thenmalarchelvi, R.: New insights into DNA triplexes: residual twist and radial difference as measures of base triplet non-isomorphism and their implication to sequence-dependent non-uniform DNA triplex. *Nucleic Acids Research.* vol. 33, no. 1. January 2005: pp. 43–55. doi:10.1093/nar/gki143.
Retrieved from: `https://doi.org/10.1093/nar/gki143`

[204] Treepong, P.; Guyeux, C.; Meunier, A.; et al.: panISa: Ab initio detection of insertion sequences in bacterial genomes from short read sequence data. *Bioinformatics*. June 2018. doi:10.1093/bioinformatics/bty479.
Retrieved from: `https://doi.org/10.1093/bioinformatics/bty479`

[205] Usmanova, D. R.; Bogatyreva, N. S.; Bernad, J. A.; et al.: Self-consistency test reveals systematic bias in programs for prediction change of stability upon mutation. *Bioinformatics*. vol. 34, no. 21. May 2018: pp. 3653–3658. doi:10.1093/bioinformatics/bty340.
Retrieved from: `https://doi.org/10.1093/bioinformatics/bty340`

[206] Vanacek, P.; Sebestova, E.; Babkova, P.; et al.: Exploration of Enzyme Diversity by Integrating Bioinformatics with Expression Analysis and Biochemical Characterization. *ACS Catalysis*. vol. 8, no. 3. February 2018: pp. 2402–2412. doi:10.1021/acscatal.7b03523.
Retrieved from: `https://doi.org/10.1021/acscatal.7b03523`

[207] Vandecraen, J.; Chandler, M.; Aertsen, A.; et al.: The impact of insertion sequences on bacterial genome plasticity and adaptability. *Critical Reviews in Microbiology*. vol. 43, no. 6. April 2017: pp. 709–730. doi:10.1080/1040841x.2017.1303661.
Retrieved from: `https://doi.org/10.1080/1040841x.2017.1303661`

[208] Varani, A. M.; Siguier, P.; Gourbeyre, E.; et al.: ISsaga is an ensemble of web-based methods for high throughput identification and semi-automatic annotation of insertion sequences in prokaryotic genomes. *Genome Biology*. vol. 12, no. 3. 2011: page R30. doi:10.1186/gb-2011-12-3-r30.
Retrieved from: `https://doi.org/10.1186/gb-2011-12-3-r30`

[209] Varizhuk, A.; Ischenko, D.; Smirnov, I.; et al.: An Improved Search Algorithm to Find G-Quadruplexes in Genome Sequences. January 2014. doi:10.1101/001990.
Retrieved from: `https://doi.org/10.1101/001990`

[210] Varizhuk, A.; Ischenko, D.; Tsvetkov, V.; et al.: The expanding repertoire of G4 DNA structures. *Biochimie*. vol. 135. April 2017: pp. 54–62. doi:10.1016/j.biochi.2017.01.003.
Retrieved from: `https://doi.org/10.1016/j.biochi.2017.01.003`

[211] Wagner, A.; Lewis, C.; Bichsel, M.: A survey of bacterial insertion sequences using IScan. *Nucleic Acids Research*. vol. 35, no. 16. August 2007: pp. 5284–5293. doi:10.1093/nar/gkm597.
Retrieved from: `https://doi.org/10.1093/nar/gkm597`

[212] Wainreb, G.; Wolf, L.; Ashkenazy, H.; et al.: Protein stability: a single recorded mutation aids in predicting the effects of other mutations in the same amino acid site. *Bioinformatics*. vol. 27, no. 23. October 2011: pp. 3286–3292. doi:10.1093/bioinformatics/btr576.
Retrieved from: `https://doi.org/10.1093/bioinformatics/btr576`

[213] Walsh, I.; Martin, A. J. M.; Domenico, T. D.; et al.: ESpritz: accurate and fast prediction of protein disorder. *Bioinformatics*. vol. 28, no. 4. December 2011: pp. 503–509. doi:10.1093/bioinformatics/btr682.
Retrieved from: `https://doi.org/10.1093/bioinformatics/btr682`

[214] Wang, G.; Vasquez, K. M.: Naturally occurring H-DNA-forming sequences are mutagenic in mammalian cells. *Proceedings of the National Academy of Sciences.* vol. 101, no. 37. September 2004: pp. 13448–13453. doi:10.1073/pnas.0405116101. Retrieved from: `https://doi.org/10.1073/pnas.0405116101`

[215] Webb, B.; Sali, A.: Protein Structure Modeling with MODELLER. In *Methods in Molecular Biology.* Springer New York. 2014. pp. 1–15. doi:10.1007/978-1-4939-0366-5_1. Retrieved from: `https://doi.org/10.1007/978-1-4939-0366-5_1`

[216] Westesson, O.; Barquist, L.; Holmes, I.: HandAlign: Bayesian multiple sequence alignment, phylogeny and ancestral reconstruction. *Bioinformatics.* vol. 28, no. 8. January 2012: pp. 1170–1171. doi:10.1093/bioinformatics/bts058. Retrieved from: `https://doi.org/10.1093/bioinformatics/bts058`

[217] Wijma, H. J.; Floor, R. J.; Janssen, D. B.: Structure- and sequence-analysis inspired engineering of proteins for enhanced thermostability. *Current Opinion in Structural Biology.* vol. 23, no. 4. August 2013: pp. 588–594. doi:10.1016/j.sbi.2013.04.008. Retrieved from: `https://doi.org/10.1016/j.sbi.2013.04.008`

[218] Wijma, H. J.; Floor, R. J.; Jekel, P. A.; et al.: Computationally designed libraries for rapid enzyme stabilization. *Protein Engineering Design and Selection.* vol. 27, no. 2. January 2014: pp. 49–58. doi:10.1093/protein/gzt061. Retrieved from: `https://doi.org/10.1093/protein/gzt061`

[219] Wilkinson, D. L.; Harrison, R. G.: Predicting the Solubility of Recombinant Proteins in Escherichia coli. *Nature Biotechnology.* vol. 9, no. 5. May 1991: pp. 443–448. doi:10.1038/nbt0591-443. Retrieved from: `https://doi.org/10.1038/nbt0591-443`

[220] Witvliet, D. K.; Strokach, A.; Giraldo-Forero, A. F.; et al.: ELASPIC web-server: proteome-wide structure-based prediction of mutation effects on protein stability and binding affinity. *Bioinformatics.* vol. 32, no. 10. January 2016: pp. 1589–1591. doi:10.1093/bioinformatics/btw031. Retrieved from: `https://doi.org/10.1093/bioinformatics/btw031`

[221] Xie, Z.; Tang, H.: ISEScan: automated identification of insertion sequence elements in prokaryotic genomes. *Bioinformatics.* vol. 33, no. 21. July 2017: pp. 3340–3347. doi:10.1093/bioinformatics/btx433. Retrieved from: `https://doi.org/10.1093/bioinformatics/btx433`

[222] Xie, Z.-R.; Hwang, M.-J.: Methods for Predicting Protein–Ligand Binding Sites. In *Methods in Molecular Biology.* Springer New York. September 2014. pp. 383–398. doi:10.1007/978-1-4939-1465-4_17. Retrieved from: `https://doi.org/10.1007/978-1-4939-1465-4_17`

[223] XODO, L. E.; ALUNNI-FABBRONI, M.; MANZINI, G.; et al.: Sequence-specific DNA-triplex formation at imperfect homopurine-homopyrimidine sequences within a DNA plasmid. *European Journal of Biochemistry.* vol. 212, no. 2. March 1993: pp. 395–401. doi:10.1111/j.1432-1033.1993.tb17674.x. Retrieved from: `https://doi.org/10.1111/j.1432-1033.1993.tb17674.x`

[224] Xu, Z.; Wang, H.: LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Research.* vol. 35, no. Web Server. May 2007: pp. W265–W268. doi:10.1093/nar/gkm286.
Retrieved from: `https://doi.org/10.1093/nar/gkm286`

[225] Yang, J.; Yan, R.; Roy, A.; et al.: The I-TASSER Suite: protein structure and function prediction. *Nature Methods.* vol. 12, no. 1. December 2014: pp. 7–8. doi:10.1038/nmeth.3213.
Retrieved from: `https://doi.org/10.1038/nmeth.3213`

[226] Ye, C.; Ji, G.; Li, L.; et al.: detectIR: A Novel Program for Detecting Perfect and Imperfect Inverted Repeats Using Complex Numbers and Vector Calculation. *PLoS ONE.* vol. 9, no. 11. November 2014: page e113349. doi:10.1371/journal.pone.0113349.
Retrieved from: `https://doi.org/10.1371/journal.pone.0113349`

[227] Yin, S.; Ding, F.; Dokholyan, N. V.: Eris: an automated estimator of protein stability. *Nature Methods.* vol. 4, no. 6. June 2007: pp. 466–467. doi:10.1038/nmeth0607-466.
Retrieved from: `https://doi.org/10.1038/nmeth0607-466`

[228] Yu, H.; Dalby, P. A.: Exploiting correlated molecular-dynamics networks to counteract enzyme activity–stability trade-off. *Proceedings of the National Academy of Sciences.* vol. 115, no. 52. December 2018. doi:10.1073/pnas.1812204115.
Retrieved from: `https://doi.org/10.1073/pnas.1812204115`

[229] Yu, H.; Huang, H.: Engineering proteins for thermostability through rigidifying flexible sites. *Biotechnology Advances.* vol. 32, no. 2. March 2014: pp. 308–315. doi:10.1016/j.biotechadv.2013.10.012.
Retrieved from: `https://doi.org/10.1016/j.biotechadv.2013.10.012`

[230] Yuan, Y.; Pei, J.; Lai, L.: Binding Site Detection and Druggability Prediction of Protein Targets for Structure- Based Drug Design. *Current Pharmaceutical Design.* vol. 19, no. 12. February 2013: pp. 2326–2333. doi:10.2174/1381612811319120019.
Retrieved from: `https://doi.org/10.2174/1381612811319120019`

[231] Zerbino, D. R.: Using the Velvet de novo Assembler for Short-Read Sequencing Technologies. *Current Protocols in Bioinformatics.* vol. 31, no. 1. September 2010. doi:10.1002/0471250953.bi1105s31.
Retrieved from: `https://doi.org/10.1002/0471250953.bi1105s31`

[232] Zhang, Z.; Li, Y.; Lin, B.; et al.: Identification of cavities on protein surface using multiple computational approaches for drug binding site prediction. *Bioinformatics.* vol. 27, no. 15. June 2011: pp. 2083–2088. doi:10.1093/bioinformatics/btr331.
Retrieved from: `https://doi.org/10.1093/bioinformatics/btr331`

[233] Zhao, J.; Bacolla, A.; Wang, G.; et al.: Non-B DNA structure-induced genetic instability and evolution. *Cellular and Molecular Life Sciences.* vol. 67, no. 1. September 2009: pp. 43–62. doi:10.1007/s00018-009-0131-2.
Retrieved from: `https://doi.org/10.1007/s00018-009-0131-2`

[234] Zhao, Y.; Tang, H.; Ye, Y.: RAPSearch2: a fast and memory-efficient protein similarity search tool for next-generation sequencing data. *Bioinformatics.* vol. 28, no. 1. October 2011: pp. 125–126. doi:10.1093/bioinformatics/btr595.
Retrieved from: `https://doi.org/10.1093/bioinformatics/btr595`

[235] Zuker, M.: Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Research.* vol. 31, no. 13. July 2003: pp. 3406–3415. doi:10.1093/nar/gkg595.
Retrieved from: `https://doi.org/10.1093/nar/gkg595`

[236] Zytnicki, M.; Akhunov, E.; Quesneville, H.: Tedna: a transposable element de novo assembler. *Bioinformatics.* vol. 30, no. 18. June 2014: pp. 2656–2658. doi:10.1093/bioinformatics/btu365.
Retrieved from: `https://doi.org/10.1093/bioinformatics/btu365`

Appendix A

Included Papers

A.1 Paper I

A dynamic programming algorithm for identification of triplex-forming sequences

BIOINFORMATICS ORIGINAL PAPER

Vol. 27 no. 18 2011, pages 2510–2517
doi:10.1093/bioinformatics/btr439

Sequence analysis

Advance Access publication July 16, 2011

A dynamic programming algorithm for identification of triplex-forming sequences

Matej Lexa[1,*], Tomáš Martínek[2], Ivana Burgetová[2], Daniel Kopeček[1] and Marie Brázdová[3]

[1]Department of Information Technology, Faculty of Informatics, Masaryk University, 60200 Brno, [2]Department of Computer Systems, Faculty of Information Technology, Brno Technical University, 61266 Brno and [3]Department of Biophysical Chemistry and Molecular Oncology , Institute of Biophysics, Academy of Sciences of the Czech Republic v.v.i., CZ-61265 Brno, Czech Republic

Associate Editor: John Quackenbush

ABSTRACT

Motivation: Current methods for identification of potential triplex-forming sequences in genomes and similar sequence sets rely primarily on detecting homopurine and homopyrimidine tracts. Procedures capable of detecting sequences supporting imperfect, but structurally feasible intramolecular triplex structures are needed for better sequence analysis.

Results: We modified an algorithm for detection of approximate palindromes, so as to account for the special nature of triplex DNA structures. From available literature, we conclude that approximate triplexes tolerate two classes of errors. One, analogical to mismatches in duplex DNA, involves nucleotides in triplets that do not readily form Hoogsteen bonds. The other class involves geometrically incompatible neighboring triplets hindering proper alignment of strands for optimal hydrogen bonding and stacking. We tested the statistical properties of the algorithm, as well as its correctness when confronted with known triplex sequences. The proposed algorithm satisfactorily detects sequences with intramolecular triplex-forming potential. Its complexity is directly comparable to palindrome searching.

Availability: Our implementation of the algorithm is available at http://www.fi.muni.cz/~lexa/triplex as source code and a web-based search tool. The source code compiles into a library providing searching capability to other programs, as well as into a stand-alone command-line application based on this library.

Contact: lexa@fi.muni.cz

Supplementary Information: Supplementary data are available at Bioinformatics online.

Received on March 16, 2011; revised on June 24, 2011; accepted on July 13, 2011

1 INTRODUCTION

Triplexes are local structural variants of DNA, wherein the molecule adopts a specific secondary structure differing from a canonical duplex by the recruitment of a third DNA strand. The third strand binds to the duplex by Hoogsteen or reverse Hoogsteen bonds with stringency of the same order of magnitude as duplex-forming strands for the most stable nucleotide combinations (reviewed by Frank-Kamenetskii and Mirkin, 1995). Depending on the source of the third strand, triplex DNA can be *intrastrand* and *interstrand*, or *intramolecular* and *intermolecular*. The third strand may just come from the other strand of the same DNA duplex or from a completely different DNA molecule, as is the case with triplex-forming oligonucleotides (Knauert and Glazer, 2001). Nucleotides in the middle strand of a triplex have Watson–Crick base pairing to one nucleotide and Hoogsteen or reverse Hoogsteen pairing to another nucleotide. Together they form a triplex-forming triplet (also called triad) (Mirkin and Frank-Kamenetskii, 1994; Soyfer and Potaman, 1995). Depending on the orientation of the third strand, we distinguish *parallel* and *antiparallel* triplexes, named according to the orientation of the third strand in respect to the central strand. Figure 1 shows eight types of *intramolecular* triplex structures considered in this article. A given sequence on the (+) strand of a DNA molecule can possibly support all eight types, but necessarily, only one of the types will be formed at any particular moment. In DNA triplexes, there is a requirement for neighboring triplets to be isomorphic, otherwise the potential triplet would be under strain, hindering the binding of the third strand (Rathinavelan and Yathindra, 2006; Thenmalarchelvi and Yathindra, 2005). Regardless of orientation and geometry, the middle nucleotide is generally a purine-containing one, to support the extra hydrogen bonds needed to bind the third nucleotide.

Because the middle nucleotide is almost invariably one with a purine base, attempts to correlate sequence with triplex-forming properties usually involve detection of homopurine and homopyrimidine tracts in the analyzed sequence. For example, Gaddis *et al*. (2006) created a web-based program that identifies target sequences for triplex-forming oligonucleotides. The program identifies homopurine stretches that are allowed to be occasionally interrupted by a pyrimidine. While this is an appropriate method for detection of strong triplex-forming signals, we consider this to be an oversimplification. Numerous papers have reported the existence of imperfect triplexes (Mergny *et al*., 1991; Roberts and Crothers, 1991; Xodo *et al*., 1993), including cases where the authors deliberately changed individual nucleotides to observe the effects of such change. Changes resulting in the formation of non-canonical triplets did not necessarily disrupt the entire triplex. It is conceivable that many of the imperfect triplexes may still have similar biological activity to their ideal counterparts. One possible explanation for the existence of imperfect triplexes is that they may allow an overlap between the structural signal and some other sequence feature,

[*]To whom correspondence should be addressed.

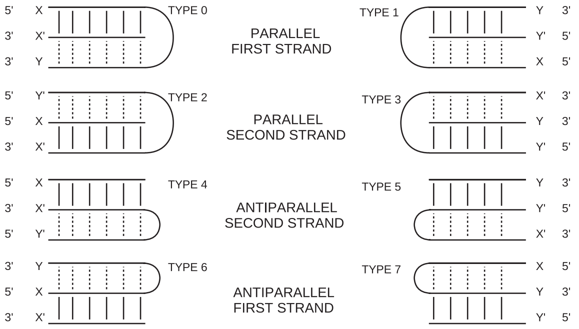


Fig. 1. Eight types of triplexes that are detected in separate runs of the algorithm for a given region. Numbering of types is shown as used in the accompanying software (Supplementary Material). Watson–Crick base pairing is shown by vertical bars. X and Y are two nucleotides on the same strand that will form a triplet. The eight possible triplets are: Y.X′X, Y′.XX′, Y′.X′X, Y.XX′, X.Y′Y, X′.YY′, X′.Y′Y and X.YY′ (N′, a nucleotide complementary to N; '.', Hoogsteen or reverse Hoogsteen bond).

such as nucleosome positioning pattern or a regulatory protein-binding sequence. Kinniburgh (1989) proposed a triplex structure containing a single deletion to explain his experimental results. Additionally, analyzed sequences may contain errors, including occasional deletions and insertions.

The existence of triplex DNA has been repeatedly associated with important biological processes at the molecular level, making them an attractive target in sequence analysis. Most of the observed associations suggest roles in mutagenesis, recombination and gene regulation. Non-B DNA structures, including DNA triplexes, have been shown to cause deletions, expansions and translocations in both prokaryotes and eukaryotes (Raghavan *et al.*, 2005). Their distribution is not random and often colocalizes with sites of chromosomal breakage (Zhao *et al.*, 2010). Triplex structures can block the replication fork and result in double-stranded breaks (Dixon *et al.*, 2008). Unlike other non-canonical structures, triplex-forming sequences are found frequently in promoters and exons and have been found to be involved in regulating the expression of several disease-linked genes (Wang and Vasquez, 2004). In some cases, the mutagenesis induced by such sequences is enhanced by their transcription (Belotserkovskii *et al.*, 2007), possibly via transcriptional arrest.

Sequence–structure relationships of triplexes were brought into a small number of computational tools for identifying relevant sequences in genome sequences. Schroth and Ho (1995) analyzed the occurrence of inverted and mirror repeats in three genomes. Hoyne *et al.* (2000) analyzed the *Escherichia coli* genome for intrastrand triplex sequences. Another recent work (Cer *et al.*, 2010) created a web-based catalog of non-B DNA sequences in major mammalian genomes. Their definition of triplex covers the most stable canonical triplexes made of G.GC/A.AT and C.GC/T.AT triplets, but leaves little room for possible errors. Jenjaroenpun and Kuznetsov (2009) created a web-based analysis tool for triplex target sequences.

Intramolecular triplex DNA (also called H-DNA) has been shown to exist both *in vivo* and *in vitro* (Hanvey *et al.*, 1988). Its formation also depends on the topological state of the given DNA molecule. While sequences supporting canonical triplets, such as $(CT(T))_n$ and $(GA(A))_n$ tracts, form triplexes readily, imperfect triplexes may require special conditions, such as low superhelical density or certain pH to form. *In vitro*, superhelical density and pH can be easily controlled. *In vivo*, pH is tightly controlled by the cell, while the topological state of any stretch of genomic DNA is generally unknown, but presumed to be under regulatory control as well. This uncertainty is the main reason for using the term 'triplex-forming sequence' or 'triplex-forming potential', which hints that while the sequence should be capable of forming a triplex, it may only be formed under special circumstances.

2 APPROACH

Based on available literature, we assume there are two important classes of sequence-based imperfections (errors) destabilizing potential triplex structures.

- Base pairing mismatch
- Geometrical mismatch

A base pairing mismatch occurs upon the formation of a nucleotide triplet that does not support strong Hoogsteen or reverse Hoogsteen bonds. The ability to form the bond and its strength is related to the number of hydrogen bonds that can be made between the second and third strand base. In this article, we present an algorithm that is based on scores assigned to base triplets. The scores are meant to approximate energy contributions of individual triplets, but at the same time to be simple enough to support rapid searching that could be used as pre-filtering, preceding detailed energy calculations on the candidate sequences.

A geometrical mismatch occurs when directly neighboring triplets in a structure are not isomorphic. This places extra stress on the backbone of the third DNA strand preventing it from creating optimal hydrogen bonds. According to Thenmalarchelvi and Yathindra (2005), conformational changes necessitated by triplet non-isomorphism are found to induce an alternative zig-zag backbone structure for the third strand in special cases. Accordingly, we made our algorithm favor triplet combinations that are either isomorphic or made of non-isomorphic pairs that could form a zig-zag shape by canceling their geometric effect on the third strand backbone.

We currently ignore other known factors of triplex DNA formation, such as the competition between alternative structures (Rippe *et al.*, 1992), fourth strand (the strand which is not part of the predicted triplex) secondary structure, effects of C+ distribution (James *et al.*, 2003; Seidman and Glazer, 2003) and other distortions caused by electrostatic forces (Kang *et al.*, 1992; Tan and Chen, 2006). Most of these factors depend non-trivially on the environment (Plum *et al.*, 1995). Since the algorithm does not consider the environment, we focus primarily on sequence-coded effects and the resulting constraints which can be computed using the information from primary structure. Destabilizing effects of loop lengths that differ from the optimum of about five nucleotides (Haasnoot *et al.*, 1986) and the overall length of the triplex (Tan and Chen, 2006) are partly accounted for, since these parameters can be set as hard limits in our implementation, to narrow the search space.

3 METHODS

Datasets: to evaluate the algorithm on selected datasets, we prepared a set of sequences to work with (all ~4.7 Mb to match the size of *E.coli* genome): (i) a random nucleotide sequence; (ii) *E.coli* K-12 MG1655 complete genome (the 1995 U00096.1 version to be able to compare our results to previous publications); (iii) *E.coli* K-12 MG1655 complete genome (the current U00096.2 version for proper positioning in genome browsers); (iv) a randomized nucleotide sequence of the same *E.coli* genome; (v) a part of the human chromosome 5 sequence (positions 144635154–149340649) and (vi) a randomized version of the same human sequence. For the human randomized sequence, we also generated a triplex-seeded version with 418 triplex-forming sequences from literature inserted at positions ∼10 000 bp apart. All the sequence data are available as Supplementary Material and can also be downloaded from *http://www.fi.muni.cz/∼lexa/triplex*. Random sequences were generated with equal probability for all four bases, and were prepared with an in-house algorithm seqmix-0.2 (Supplementary Material).

Molecular simulations of triplets: to obtain objective information about isomorphic groups, we analyzed the angle and radius formed by C1 atoms of triplet nucleotides as defined in Thenmalarchelvi and Yathindra (2005). The groups were determined using the following procedure. First, the structures of all considered triplets were constructed using the NAB language from AmberTools 1.4 and their potential energy surface was explored for local minima by moving and rotating the third (Hoogsteen) base in the plane formed by the other two bases. The energy function was parametrized using the *ff99bsc0* set (Perez *et al*., 2007). The obtained local minima were filtered according to the values of the C1 angle (t) and the ratio $|WH|/|CH|$, where $|WH|$ represents the distance between the C1 atoms of the Hoogsteen pair and $|CH|$ represents the distance between the C1 atoms of the mutually unpaired bases. Filtering thresholds were derived from measurements on a set of real structures, namely the structures 135D, 149D, 1BCB, 1D3X (PDB identifiers). The specific thresholds used were $70 \leq t \leq 130$ and $0.54 \leq |WH|/|CH| \leq 0.88$. From the resulting set of local minima, the structure with the lowest potential energy was selected as the source of the parameters t and r (the radius of the circle formed by the C1 atoms). Finally, the groups were established by performing cluster analysis using Ward's method and euclidean distance between the (t,r) vectors. These results were interpreted to obtain isomorphic groups in Table 1, and detailed results are available as Supplementary Material.

Testing overview: we tested our implementation for correctness and usability. Clearly, the algorithm will only be useful, if it is capable of identifying potential triplex-forming sequences in a genomic background with a reasonable success rate. To test the implementation in this respect, we performed statistical tests on real and randomized sequences, a sequence recovery test on the triplex-seeded sequences, and we compared our solution to previously published results for the *E.coli* genome (Hoyne *et al*., 2000) and a currently published non-B DNA database (Cer *et al*., 2010).

Statistical tests: the statistical tests served to find parameters for the distribution of scores on randomized sequences and establish a proper threshold above which candidate hits should be considered significant. The distribution of scores was modeled according to principles used for evaluating BLAST results and other sequence similarity scores (Altschul *et al*., 1994; Korf *et al*., 2003), since the alignment of a DNA strand against itself is statistically similar to aligning two different sequences. This treatment allowed us to fit the score distribution with an extreme value distribution function and fit the parameters λ and μ as described by Korf *et al*. (2003). To carry out the calculation, we used a function from hmmer-2.3.2 source code (Eddy, 1997).

Recovery tests: the recovery tests evaluated how many of the introduced triplex-forming sequences were recovered for a selected significance threshold (P-value) from different backgrounds sequences. We used the commonly used characteristics for such experiments: specificity (precision), sensitivity (recall), F_2 measure and accuracy (Manning *et al*., 2008). The algorithm was tested against our triplex-seeded sequence and a database of non-B DNA (Cer *et al*., 2010).

Table 1. Triplex scoring of canonical and less usual triplets

Triplex type	Triplet H.WC:WC	Score (*tts*)	Isomorphic group	References
PARALLEL	T.A:T	2	a	Goni *et al*. (2004)
	T.G:C	1	a	Ghosal and Muniyappa (2006)
	C.G:C	2	a	Walter *et al*. (2001),Goni *et al*. (2004)
	G.G:C	1	b	Soyfer and Potaman (1995)
	G.T:A	2	b	Gowers and Fox (1998)
	T.C:G	1	b	Soyfer and Potaman (1995)
ANTIPARALLEL	A.A:T	2	c	(Goni *et al*. (2004), Mirkin and Frank-Kamenetskii (1994))
	A.G:C	1	d	Mirkin and Frank-Kamenetskii (1994), Raghavan and Lieber (2007)
	T.A:T	2	c	Goni *et al*. (2004), Mirkin and Frank-Kamenetskii (1994)
	T.C:G	1	e	Raghavan and Lieber (2007), Beal and Dervan (1992)
	C.A:T	1	d	Raghavan and Lieber (2007), Soyfer and Potaman (1995),Dayn *et al*. (1992)
	G.G:C	2	e	Goni *et al*. (2004), Mirkin and Frank-Kamenetskii (1994)

The final score values for both Hoogsteen and reverse-Hoogsteen bonds are in accordance with tables 4.1 and 4.2 in Soyfer and Potaman (1995). Isomorphic groups shown here are based on residual twist calculations using molecular dynamics simulations with the *nbd* program (AmberTools). ., Hoogsteen bp; :, Watson–Crick bp; *tts*, tabulated triplet score.

Escherichia coli tests: we compared our tool and its performance on the *E.coli* genome sequence to the results published by Hoyne *et al*. (2000). Additionally, we calculated the genome positioning of program output in respect to known *E.coli* genes, counting the frequency with which predicted triplexes fell inside the gene, outside any genes or intersected with them. Distance to the closest gene was calculated as shown in Figure 6.

4 THE ALGORITHM

Our approach to search for approximate triplexes is based on a dynamic programming (DP) algorithm to search for approximate palindromes that can be traced back to Landau and Vishkin (1986). The relationship between triplex DNA and palindromes stems from the fact, that one of the DNA strands in the triplex must fold back onto itself, either for Hoogsteen base pairing or for reverse Hoogsteen base pairing, depending on the type of triplex that is to be formed (parallel or antiparallel) and the nucleotide sequence present at the site in question. We will call the part of the triplex that folds back onto itself *self-recognizing*.

A DP matrix is constructed so that one side represents the original sequence, while the other contains the same sequence written backwards (Fig. 2). With such setup, the main antidiagonal of the DP matrix represents the n possible central starting positions for the self-recognizing parts of triplexes with an odd number of nucleotides in the loop. The neighboring antidiagonal contains the other $n-1$ possible starting sites for the triplexes with even number of nucleotides in their loops. Naturally, diagonals starting at any of these positions represent potential triplexes. If we fill the cells representing the starting positions with zeros, we can start filling the DP matrix along the diagonals. At each position $[i,j]$ of the DP matrix, we compare the symbols at positions i and j in the original sequence. If they represent a pair present in triplex-forming triplets (tabulated in Table 1), they are evaluated with positive score. In opposite case, they are penalized with a negative score value. The numbers entered represent the best score in the subsequence evaluated so far.

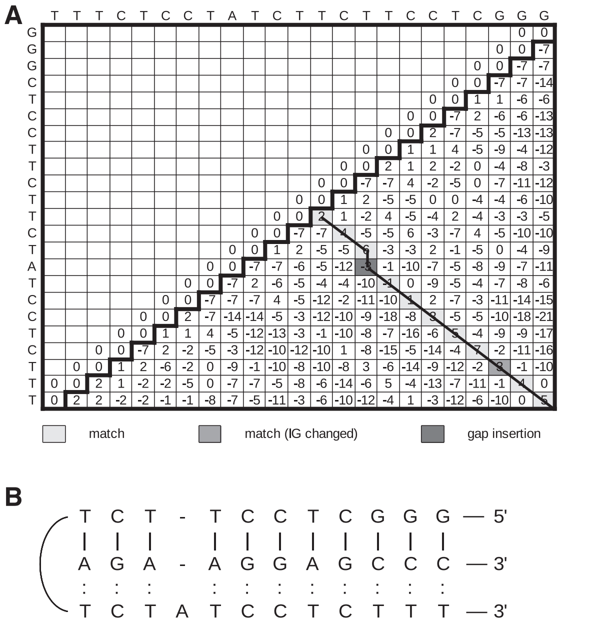

A
T T T C T C C T A T C T T C T T C C T C G G G
G G G C T C C T T C T T C T A T C C T C T T T
match
match (IG changed)
gap insertion
B
T C T - T C C T C G G G — 5'
A G A - A G G A G C C C — 3'
T C T A T C C T C T T T — 3'

Fig. 2. Triplex detection by the DP algorithm demonstrated on the string *gggctccttcttctatcctcttt*. (**A**) The DP matrix with calculated score values. Because of space limitations, loop size was forced to 0. (**B**) Triplex alignment. Hoogsteen bonds are shown by semicolons.

The necessity for a dynamic programming algorithm comes from the possibility to insert gaps into the triplexes, where symbols in some positions have no symbols to pair up with in the other arm of the self-recognizing sequence. In terms of the described algorithm, this means moving from one diagonal to a neighboring one when calculating the score. At any position, three possibilities are evaluated:

(1) Extending the existing triplex along the diagonal - *match* or *mismatch*,

(2) Inserting a gap at position *i* of the original sequence - *insertion*,

(3) Inserting a gap at position *j* of the original sequence - *deletion*.

The solution that leads to the maximum score value is recorded in the DP matrix, while the other possibilities are discarded.

In comparison to a similar algorithm for approximate palindrome detection, we have introduced three important modifications. First, we redefined the concept of match and mismatch. Instead of being made up by pairs of nucleotides with only two possible base pairs, triplexes can be thought of as sequences of triplets with many possible combinations of nucleotides in the triplet. There are 16 possible base pairs for parallel DNA strands and another 16 for antiparallel strands. For these reasons, we constructed a general similarity matrix instead of using a single match rule and score.

Second modification brings geometrical considerations into the algorithm, making certain sequences of triplets less desirable than others. This is similar to the nearest-neighbor scoring used in duplexes, although we are not as much concerned about base stacking as we are about the geometry of the third strand and its ability to position itself for optimal hydrogen bonding. As discussed by Rathinavelan and Yathindra (2006); Thenmalarchelvi and Yathindra (2005), some combinations disrupt the backbone geometry. We therefore decided to divide the triplets into isomorphic groups. Groups of triplets from one group are more likely to form stable triplexes than other sequences. Our modification assigns the information about isomorphic groups to the last computed DP matrix cell on each diagonal. When calculating a new cell, we lower the score if the newly evaluated triplet belongs to a different isomorphic group than the preceding one. The score calculation is

$$S[i,j]=\max\begin{cases}S[i,j-1]+gp\\ S[i-1,j]+gp\\ S[i-1,j-1]+tts[a,b]+nip\end{cases} \qquad (1)$$

where a, b are characters at appropriate row and column, tss is tabulated triplet score, gp is gap penalty and nip is no-isomorphism penalty.

The third consideration is to account for all the possible ways a triplex can form from a given sequence, i.e. which three strands combine together and in which orientation (Fig. 1). There are always eight ways that can give rise to a intramolecular triplex at a given position, since there are two strands that can serve as the third strand, each having two ends that can loop back onto the double-stranded region and in each of these cases it can attach on either side of the duplex in a parallel or antiparallel fashion, forming Hoogsteen and reverse Hoogsteen bonds, respectively. In order to detect all types of triplexes the computation is repeated eight times with scoring matrices specific for parallel and antiparallel triplexes.

4.1 Scoring function

We evaluate the combinations based on their ability to form Hoogsteen base pairs, tabulating the 32 values as complementarity scores. One way to populate such table is to consider all canonical triplets to represent a match and everything else a mismatch. Because the ability to form Hoogsteen bonds depends partly on the environment of the given nucleotide, we took a semi-empirical approach, giving all canonical triplets a match score of 2, scanning triplex literature for examples of less usual triplets and giving those a score of 1, while all other combinations are scored as a mismatch (Table 1). Other approaches leading to a better scoring scheme are certainly possible, but beyond the scope of this article.

4.2 Triplex loop detection

The algorithm introduced in this section has been designed to detect the best candidates for triplex formation. To avoid the inclusion of free-strand and loop nucleotides into the overall score for a particular triplex (because these nucleotides do not participate in Watson–Crick or Hoogsteen base pairing), our calculations use a technique composed of a combination of local and global alignment.

In terms of the DP matrix, potential loops always begin at the main antidiagonal, extending up to $l_{loop_{max}}$ (user-defined algorithm parameter), using Equation (1) to calculate new values. The first $2l_{loop_{max}}$ antidiagonals are therefore calculated by a technique similar to the one used in Smith–Waterman local sequence alignment. In this part, we allow the score of a growing triplex to grow or decline. However, if the density of errors is high enough to bring the score into the negative territory (potential loop occurrence), we do not allow the score to become negative.

Once the calculations exit the area of a potential loop, the calculations continue in a global alignment mode. This way the algorithm can detect high-quality triplex candidates without considering errors that fall within potential loops.

4.3 Triplex detection

The best triplexes in the DP matrix can be identified as those reaching the highest score. To allow detection of such *high scoring segments* (HSS) during the calculation, we use a technique similar to the one used in the BLAST program. Once the score rises above a preset threshold value, the region responsible for the score is considered a potential triplex. The score is monitored (allowed both to increase and decrease) until it falls below a preset threshold. The sequence from the beginning (the first antidiagonal) up to the maximum score becomes the HSS of the potential triplex (Fig. 3).

A number of filtration mechanisms can be applied to the step of HSS segment detection. One of the problems we had to deal with (causing false HSS detection), was the transfer of scores from neighboring diagonals.

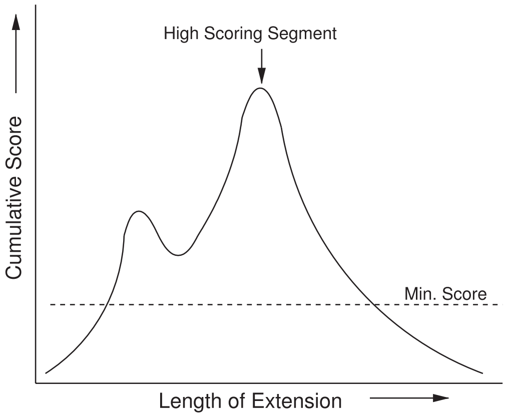


Fig. 3. Detection of high scoring segments.

In the presence of a high-quality triplex sequence, neighboring diagonals adopt its high score by introduction of an extra insertion or deletion. We therefore check for such cases and only report genuine HSS scores and not the neighboring derivatives.

Further filtration is carried out based on statistical significance of the results, eliminating all short or low-quality potential triplexes below a user-defined E-value or P-value threshold (see Section 5 for details on P-value calculations on experimental datasets). A pair of filtering programs (prefilter_gff.c and filter_gff.c, see Supplementary Material) were used to filter out results not supporting a local score maximum (meaning there is a better result nearby).

4.4 Time and space complexity

Time complexity: the calculation of the entire triangle of the DP matrix has $n^2/2$ steps. However, when analyzing real or random sequences, the likelihood of finding a potential triplex decreases with its length (see section 5 for a detailed description of this effect). Therefore, for most practical purposes we only need to evaluate a limited number of antidiagonals, say $2l$, where l is the maximal length of detected triplexes. Time complexity thus becomes $O(2ln)$.

Space complexity: with respect to data dependencies, only the values for the last two antidiagonals are necessary for calculation. Thus, the space complexity of our algorithm is $O(2n)$.

Both simplifications/efficiency enhancements used to derive the time and space complexities allow us to easily extend the algorithm to perform an *incremental calculation*. If upon completion of the calculation we find that the number of antidiagonals was not sufficient, leaving several potential triplexes unresolved, we can pick up the score values from the last two diagonals and continue in the calculations in another $2l$ antidiagonals.

5 RESULTS AND DISCUSSION

We subjected the algorithm to increasing levels of scrutiny to verify the validity of our searching procedures, fine-tune some of the parameters and establish the biological relevance of selected results.

Initial experiments were directed towards establishing reasonable mismatch and insertion/deletion penalties. The penalties have to be high enough to allow for a negative average score per triplet (Korf *et al.*, 2003). Without any rigorous optimization, we found the combination *mismatch* −7, *insertion* or *deletion* −9, *no_isomorphism* −5 to fulfill these criteria and work reasonably well on all sequences.

Table 2. The results of fitting an extreme value distribution function to score distribution data obtained from randomized sequences of *E.coli* and human genomes

Randomized sequence data	λ	μ	Threshold
Escherichia coli	0.91	6.00	20
Human chr5	0.84	6.28	21

The threshold shown here for reference purposes is the score above which <10 sequences were found in randomized data. Precise E-values and P-values can be calculated from values of λ and μ according to Equation (2).

Identification of a higher number of potential triplexes in real-world sequences compared with random and randomized sequences is the first confirmation that the patterns we are collecting using this approach are not random, but rather specific combinations with a possible function that are less frequent in random sequences.

For a rigorous test of non-randomness of the identified candidates, we tested our implementation of the algorithm against a set of 4.7 Mb DNA sequences from *E.coli* and human genomes, their randomized version and a triplex-seeded randomized *E.coli* genome (see Section 3). For each of the sequences, we used the program to identify all potential triplexes and their scores. Since an incrementally detected triplex-forming sequence must obey similar rules as an incrementally growing sequence alignment (only with different base pairing rules), we would expect the obtained scores to obey an extreme value distribution described by Altschul *et al.* (1994).

$$P(S>x)=1-e^{-e^{-\lambda\times(x-\mu)}} \qquad (2)$$

We used a maximum likelihood method described by Eddy (1997) to fit our scores to this function. The resulting values of λ and μ are given in Table 2. Figure 4 shows a graphical representation and corresponding parameter values of triplex scores for the different datasets used. Clearly, randomized sequences have a lower content of high-scoring sequence patterns. Also, human sequences seem to be richer in potential triplex-forming sequences, comparable in density to the artificially seeded *E.coli* sequence with one triplex sequence per every 10 000 bp.

We used the λ and μ values to derive statistical thresholds for searching (Table 2). These are different for parallel and antiparallel triplexes, since the two use a different similarity matrix, resulting in different score distributions.

Next, we analyzed the non-B DNA database triplex predictions (Cer *et al.*, 2010) and our triplex-seeded sequence containing 418 inserted triplexes with artificial mismatches and insertions. Our program preferentially recovered the positions of known triplex sequences. Figure 5 shows sensitivity, specificity, accuracy and F_2 measure for these two sets. F measure is the harmonic mean of sensitivity and specificity. F_2 measure is its commonly used modification, which gives higher priority to recall. F_2 measure values >40% are satisfactory, given that 100% of potential triplexes are recovered with a P-value better than 0.01. Some loss of performance on triplex-seeded data is understandable, since mismatches and insertions/deletions were introduced in sequences as short as 6 bp.

One of the detected sequences, is a well-studied triplex from human metallothionein-I promoter (Bacolla and Wu, 1991). This sequence was the second highest-scoring sequence in the

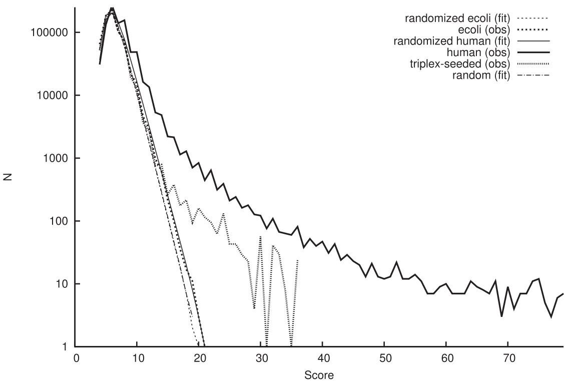


Fig. 4. Log-scale extreme value distribution functions for *E.coli* (dashed line), human (solid line) and triplex-seeded datasets (dotted line) compared with background random sequences (thin lines), including a random sequence, randomized *E.coli* and human sequences. A maximal likelihood fit to the random sequences is available in Table 2. While the *E.coli* genome contains potential triplex sequences only slightly above background levels, the human genome seems to be rich in such sequences with density similar to the triplex-seeded dataset.

triplex-seeded data, scoring 34 with a *P*-value of 5.10^{-9}. Interestingly, we detected two high-scoring subsequences within the MT-I promoter potential triplex, supporting the view of Bacolla and Wu (1991) and Becker and Maher (1998) that alternative triplex structures may be formed at this specific site.

For an alternative evaluation of the validity of our algorithm, we analyzed the *E.coli* genome for triplex-forming sequences and compared the results with those described in Hoyne *et al.* (2000). They searched for potential intrastrand triplex (PIsT). The PIsT element requires the consecutive occurrence of all three triplex-forming blocks of nucleotides, while potential intramolecular triplex (PImT) element requires the consecutive occurrence of just two triplex-forming blocks (the third block is provided by the parallel strand). Thus, every PIsT element by definition contains also a PImT element.

For each of the 25 PIsT elements presented in Hoyne *et al.* (2000), we are able to identify the corresponding PImT element in *E.coli* genome with appropriate parameter settings. The score of these elements range from the value of 6 to the value of 20 and the corresponding *P*-values vary from 4.7×10^{-1} to 2.9×10^{-6}. The best potential triplex element in *E.coli* genome found by our algorithm scored 21 with a *P*-value of 1.2×10^{-6}.

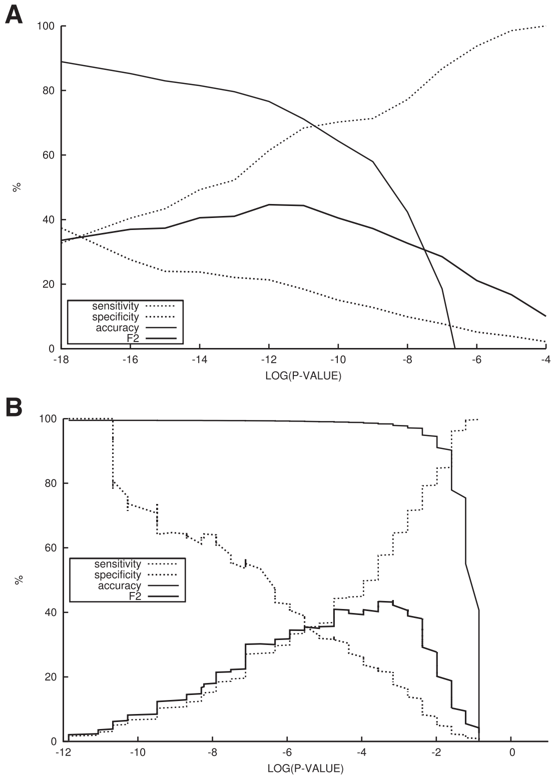


Fig. 5. Sensitivity, specificity, accuracy and F_2 measure calculated for (**A**) the non-B DNA database (Cer *et al.*, 2010); (**B**) the triplex-seeded dataset. The figure shows that the best matches obtained with the described algorithm and settings are entirely made up of the seeded sequences. At lower *P*-values, we start picking up some sequences from the background sequence; acceptable results before accuracy drops sharply are achieved for *P*-values of $<1.0\times10^{-2}$.

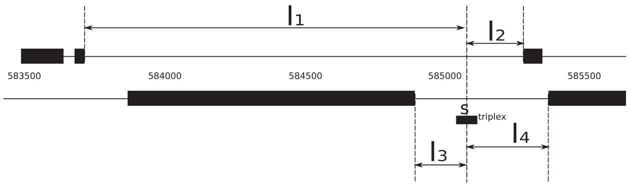


Fig. 6. The definition of the closest gene as used in the numerical experiment. For each triplex we identified its center *S* (rounded up for even triplexes), and calculated the distances l_1, l_2, l_3 and l_4 to the closest upstream and downstream gene borders on both DNA strands. The minimum of these four values was used.

Finally, we examined some of the identified potential triplex sites for biological relevance. Producing a GFF file with results enabled us to view them in the UCSC Genome Browser. Here, we noticed a possible relationship to known *E.coli* genes. To test this, we counted the number of predicted triplexes falling within genes, outside genes or <100 bp from gene boundaries (Fig. 7A). We also calculated the number of predicted triplexes occurring at different distances from the closest gene (Fig. 6) and calculated the ratio of this value to randomly placed positions. There seems to be some preference for potential triplexes to occur in the −50 to −160 region of known genes (Fig. 7B). Given the relatively high *P*-value at which this effect was still visible, it is possible that it is not directly related to the presence of triplexes, but rather a result of shared sequence characteristic between triplexes and regulatory sequences, such as their underlying palindromic nature.

Another observation showed these positions to be clustered at boundaries of evolutionarily poorly conserved regions. A quick literature search revealed a possible connection. Non-B DNA structures are likely to pose a physical barrier to transcriptional apparatus, causing possible transcriptional arrest at such sites (Young *et al.*, 1991). Transcriptional arrest has been directly linked to increased mutation rate (Belotserkovskii *et al.*, 2007), which could explain some aspect of the above-mentioned positioning in genomes.

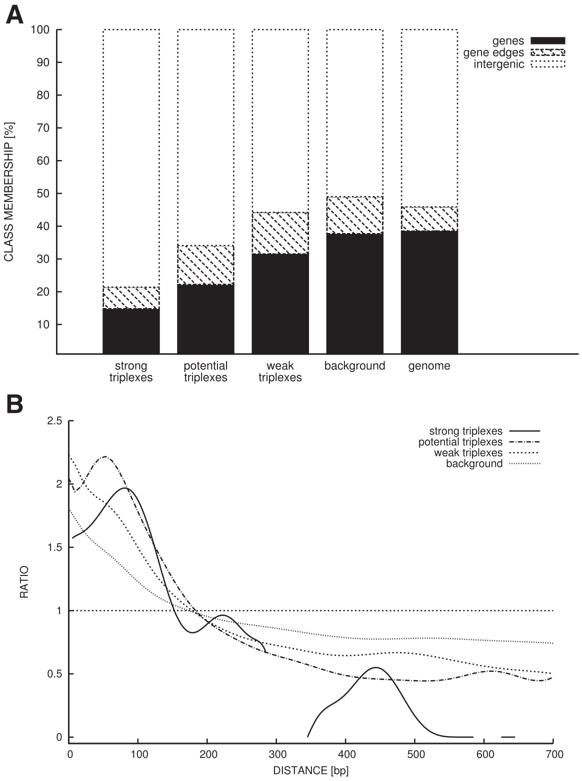


Fig. 7. Graphs showing how potential triplexes identified by the program are positioned in respect to genes in *E.coli*. (**A**) The percentage of triplexes in the results falling inside genes, intersecting with a gene or falling within intergenic segments of the genome. Bars are shown for results of decreasing specificity (from left to right); (**B**) the relative abundance of high-scoring sequences at different distances from nearby genes (relative to randomly placed positions). Both figures were generated after applying the following cutoffs to the results: top 122 (strong triplex), top 1391 (potential triplex), top 15300 (weak triplex), top 106623 (background) and random selection of positions (genome).

While the main purpose of this article is to present the algorithm itself, a more detailed analysis of the best parameter settings and performance with specific DNA sequences is needed to further increase confidence in this kind of sequence analysis.

Because of the increased complexity of scoring, the outlined procedure for scoring individual triplets within the DP matrix cannot be easily extended to take advantage of suffix arrays as is done with palindromes, to further speed up computation.

Overall, we consider it an advantage that triplex identification can be mapped to a well-researched family of DP algorithms and possibly take advantage of approaches aimed originally at other problems, such as sequence alignment.

6 CONCLUSION

We present a novel approach to identifying triplex-forming sequences in genomes and other DNA sequence data. The approach is presented in the form of an algorithm based on previously published algorithms for detection of palindromes. The novelty stems from the adaptation of DP for use with triplexes instead of relying on simpler identification of homopurine and homopyrimidine tracts, which are most appropriate for detection of perfect triplexes. We implemented our algorithm as a program written in C, using a reasonable set of parameters based on published data. The test runs of this program are encouraging, suggesting that the algorithm can provide high speed searches with increased sensitivity for approximate triplex-forming sequences.

Funding: Grants (No. 204/08/1560 and No. 301/10/2370) from the Czech Grant Agency; MSMT Research Grant no. (0021630528), Security-Oriented Research in Information Technology; BUT grants (FIT-S-11-1), Advanced secured, reliable and adaptive IT and (FIT-S-11-2), Recognition and presentation of multimedia data; ASCR Institutional Research Plans (No. AV0Z50040507 and No. AV0Z50040702).

Conflict of Interest: none declared.

REFERENCES

Altschul,S.F. *et al.* (1994) Issues in searching molecular sequence databases. *Nat. Genet.*, **6**, 119–129.

Bacolla,A. and Wu,F.Y-H. (1991) Mung bean nuclease cleavage pattern at a polypurine-polypyrimidine sequence upstream from the mouse metallothionein-I gene. *Nucleic Acids Res.*, **1**, 1639–1647.

Beal,P.A. and Dervan,P.B. (1992) The influence of single base triplet changes on the stability of a pur.pur.pyr triple helix determined by affinity cleaving. *Nucleic Acids Res.*, **20**, 2773–2776.

Becker,N.A. and Maher, L.J. III (1998) Characterization of a polypurine/polypyrimidine sequence upstream of the mouse metallothionein-I gene. *Nucleic Acids Res.*, **26**, 1951–1958.

Belotserkovskii,B.P. *et al.* (2007) A triplex-forming sequence from the human c-MYC promoter interferes with DNA transcription. *J. Biol. Chem.*, **282**, 32433–32441.

Cer,R.Z. *et al.* (2010) Non-B DB: a database of predicted non-B DNA-forming motifs in mammalian genomes. *Nucleic Acids Res.*, **39**, D383–D391.

Dayn,A. *et al.* (1992) Intramolecular DNA triplexes: unusual sequence requirements and influence on DNA polymerization. *Proc. Natl Acad. Sci. USA*, **89**, 11406–11410.

Dixon,B.P. *et al.* (2008) RecQ and RecG helicases have distinct roles in maintaining the stability of polypurine.polypyrimidine sequences. *Mutat Res.*, **643**, 20–28.

Eddy,S.R. (1997) Maximum likelihood fitting of extreme value distributions. *Technical Report*. Available at ftp://selab.janelia.org/pub/publications/Eddy97b/Eddy97b-techreport.pdf (last accessed date August 07, 2011).

Frank-Kamenetskii,M.D. and Mirkin,S.M. (1995) Triplex DNA structures. *Annu. Rev. Biochem.*, **64** 65–95.

Gaddis,S.S. *et al.* (2006) A web-based search engine for triplex-forming oligonucleotide target sequences. *Oligonucleotides*, **16**, 196–201.

Ghosal,G. and Muniyappa,P. (2006) Hoogsteen base-pairing revisited: resolving a role in normal biological processes and human diseases. *Biochem. Biophys. Res. Commun.*, **343**, 1–7.

Goni,J.R. *et al.* (2004) Triplex-forming oligonucleotide target sequences in the human genome. *Nucleic Acids Res.*, **32**, 354–360.

Gowers,D.M. and Fox,K.R. (1998) Triple helix formation at $(AT)_n$ adjacent to an oligopurine tract. *Nucleic Acids Res.*, **26**, 3626–3633.

Haasnoot,C.A.G. *et al.* (1986) On loop folding in nucleic acid hairpin-type structures. *J. Biomol. Struct. Dyn.*, **3**, 843–857.

Hanvey,J.C. *et al.* (1988) Intramolecular DNA triplexes in supercoiled plasmids. *Proc. Natl Acad. Sci. USA*, **85** 6292–6296.

Hoyne,P.R. *et al.* (2000) Searching genomes for sequences with the potential to form intrastrand triple helices. *J. Mol. Biol.*, **302**, 797–809.

James,P.L. *et al.* (2003) Thermodynamic and kinetic stability of intermolecular triple helices containing different proportions of C+·GC and T·AT triplets. *Nucleic Acids Res.*, **31**, 5598–5606.

Jenjaroenpun,P. and Kuznetsov,V.A. (2009) TTS Mapping: integrative WEB tool for analysis of triplex formation target DNA sequences, G-quadruplets and non-protein coding regulatory DNA elements in the human genome. *BMC Genomics*, **10** (Suppl. 3), S9.

Kang,S.M. *et al.* (1992) Metal ions cause the isomerization of certain intramolecular triplexes. *J. Biol. Chem.*, **267**, 1259–1264.

Kinniburgh,A.J. (1989) A cis-acting transcription element of the c-myc gene can assume an H-DNA conformation. *Nucleic Acids Res.*, **17**, 7771–7778.

Knauert,M.P. and Glazer,P.M. (2001) Triplex forming oligonucleotides: sequence-specific tools for gene targeting. *Hum. Mol. Genet.*, **10**, 2243–2251.

Korf,I. *et al.* (2003) *BLAST*. O'Reilly & Associates, Inc., Sebastopol, 368 pages.

Landau,G.M. and Vishkin,U. (1989) Fast parallel and serial approximate string matching. *J. Algorithms*, **10**, 157–169.

Manning,C.D. *et al.* (2008) *Introduction to Information Retrieval*. Cambridge University Press, Cambridge, 496 pp.

Mergny,J.L. *et al.* (1991) Sequence specificity in triple helix formation: experimental and theoretical studies of the effect of mismatches on triplex stability. *Biochemistry*, **30**, 9791–9798.

Mirkin,S.M. and Frank-Kamenetskii,M.D. (1994) H-DNA and related structures. *Annu. Rev. Biophys. Biomol. Struct.*, **23**, 541–576.

Perez,A. *et al.* (2007) Refinement of the AMBER force field for nucleic acids: improving the description of α/γ conformers. *Biophys. J.*, **92**, 3817–3829.

Plum,G.E. *et al.* (1995) Nucleic acid hybridization: triplex stability and energetics. *Annu. Rev. Biophys. Biomol. Struct.*, **24**, 319–350.

Raghavan,S.C. and Lieber,M.R. (2007) DNA structure and human diseases. *Front. Biosci.*, **12**, 4402–4408.

Raghavan,S.C. *et al.* (2005) Evidence for a triplex DNA conformation at the bcl-2 major breakpoint region of the t(14;18) translocation. *J. Biol. Chem.*, **280**, 22749–22760.

Rathinavelan,T. and Yathindra,N. (2006) Base triplet nonisomorphism strongly influences DNA triplex conformation: effect of nonisomorphic G* GC and A* AT triplets and bending of DNA triplexes. *Biopolymers*, **82**, 443–461.

Rippe,K. *et al.* (1992) Alternating d(G-A) sequences form a parallel-stranded DNA homoduplex. *EMBO J.*, **11**, 3777–3786.

Roberts,R.W. and Crothers,D.M. (1991) Specificity and stringency in DNA triplex formation. *Proc. Natl Acad. Sci. USA*, **88**, 9397–9401.

Schroth,G.P. and Ho,P.S. (1995) Occurrence of potential cruciform and H-DNA forming sequences in genomic DNA. *Nucleic Acids Res.*, **23**, 1977–1983.

Seidman,M.M. and Glazer,P.M. (2003) The potential for gene repair via triple helix formation. *J. Clin. Invest.*, **112**, 487–494.

Soyfer,V.N. and Potaman,V.N. (1995) *Triple-Helical Nucleic Acids*. Springer, Heidelberg, 360 pp.

Tan,Z.J. and Chen,S.J. (2006) Nucleic acid helix stability: effects of salt concentration, cation valence and size, and chain length. *Biophys. J.*, **90**, 1175–1190.

Thenmalarchelvi,R. and Yathindra,N. (2005) New insights into DNA triplexes: residual twist and radial difference as measures of base triplet non-isomorphism and their implication to sequence-dependent non-uniform DNA triplex. *Nucleic Acids Res.*, **33**, 43–55.

Walter,A. *et al.* (2001) Evidence for a DNA triplex in a recombination-like motif: I. Recognition of Watson-Crick base pairs by natural bases in a high-stability triplex. *J. Mol. Recognit.*, **14**, 122–139.

Wang,G. and Vasquez,K.M. (2004) Naturally occurring H-DNA-forming sequences are mutagenic in mammalian cells. *Proc. Natl Acad. Sci. USA*, **101**, 13448–13453.

Xodo,L.E. *et al.* (1993) Sequence-specific DNA-triplex formation at imperfect homopurine-homopyrimidine sequences within a DNA plasmid. *Eur. J. Biochem.*, **212**, 395–401.

Young,S.L. *et al.* (1991) Triple helix formation inhibits transcription elongation in vitro. *Proc. Natl Acad. Sci. USA*, **88**, 10023–10026.

Zhao,J. *et al.* (2010) Non-B DNA structure-induced genetic instability and evolution. *Cell Mol. Life Sci.*, **67**, 43–62.

A.2 Paper II

Triplex: an R/Bioconductor package for identification and visualization of potential intramolecular triplex patterns in DNA sequences

BIOINFORMATICS APPLICATIONS NOTE
Vol. 29 no. 15 2013, pages 1900–1901
doi:10.1093/bioinformatics/btt299

Sequence analysis

Advance Access publication May 24, 2013

Triplex: an R/Bioconductor package for identification and visualization of potential intramolecular triplex patterns in DNA sequences

Jiří Hon[1], Tomáš Martínek[1], Kamil Rajdl[2] and Matej Lexa[2,*]

[1]Department of Computer Systems, Faculty of Information Technology, Brno Technical University, Božetěchova 2, 61266 Brno and [2]Department of Information Technology, Faculty of Informatics, Masaryk University, Botanická 68a, 60200 Brno, Czech Republic

Associate Editor: Alfonso Valencia

ABSTRACT

Motivation: Upgrade and integration of triplex software into the R/Bioconductor framework.

Results: We combined a previously published implementation of a triplex DNA search algorithm with visualization to create a versatile R/Bioconductor package 'triplex'. The new package provides functions that can be used to search Bioconductor genomes and other DNA sequence data for occurrence of nucleotide patterns capable of forming intramolecular triplexes (H-DNA). Functions producing 2D and 3D diagrams of the identified triplexes allow instant visualization of the search results. Leveraging the power of Biostrings and GRanges classes, the results get fully integrated into the existing Bioconductor framework, allowing their passage to other Genome visualization and annotation packages, such as GenomeGraphs, rtracklayer or Gviz.

Availability: R package 'triplex' is available from Bioconductor (bioconductor.org).

Contact: lexa@fi.muni.cz

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on April 3, 2013; revised on May 21, 2013; accepted on May 22, 2013

1 INTRODUCTION

DNA sequence analysis and annotation are important steps in uncovering the molecular basis of life. Although protein-coding sequences have been intensively studied in the past, recent focus has shifted toward the less-known biological functions encoded in intergenic DNA, as well as the study of structural and regulatory aspects of genetic information packaging in chromosomes. Tools for the necessary sequence analysis of non-coding sequences are less common than their gene-centered counterparts. We have recently formulated and implemented an algorithm to detect potential triplex-forming sequences in genomes (Lexa *et al*., 2011). Such sequences have been implicated as important players in several key processes, such as transcriptional regulation (Walter *et al*., 2001) or DNA recombination (Rooney and Moore, 1995).

Triplex DNA forms when a third strand of nucleotides is allowed to align with a Watson–Crick duplex using Hoogsteen bonds to stabilize the nascent structure (Soyfer and Potaman, 1995). H-DNA is a form of DNA where triplexes form intramolecularly, without the participation of other DNA molecules (Htun and Dahlberg, 1989).

Currently, several research groups reported on their efforts to map triplex-forming sites in known genomes, as well as on the development of tools to carry out such searches. Hoyne *et al*. (2000) used pattern recognition tools to search for homopurine/homopyrimidine stretches in DNA as likely triplex formation sites. Cer *et al*. (2012) created a non-B DNA search tool (nBMST) that includes mirror repeat detection functionality to identify potential triplexes. Buske *et al*. (2012) and Lexa *et al*. (2011) created triplex detection procedures allowing for a small percentage of imperfections in the sequences, leading to higher sensitivity of searches. Often, the tools exist as stand-alone software or web tools, which led us to the idea to integrate triplex search, visualization and genome annotation into a unified Bioconductor software package in R for increased flexibility.

Here, we describe *triplex*, demonstrating its use in sequence analysis of sample data, focusing on functions integrating it with the rest of the R/Bioconductor suite. Of the aforementioned softwares, only *triplex* provides specialized H-DNA searching. The other software treats H-DNA as general mirror repeats and lacks fine-grained or configurable mismatch evaluation (nBMST), focuses on a different class of triplexes (Hoyne *et al*., 2000) or provides general results that need to be further filtered to identify H-DNA (triplexator), requiring several orders of processing time more than *triplex*. The software by Lexa *et al*. (2011) used to create the package was improved by (i) integration into R/Bioconductor, (ii) elimination of recognized bugs in scoring and alignment and by (iii) providing base pair information, either as text/variables or visualizations.

We performed a simple comparison of nBMST and triplexator programs with *triplex* (see Supplementary Material). It showed that reported (CT)n and (TA)n mirror repeats coincide with H-DNA found by *triplex*. Triplexator returned several longer patterns reported by *triplex* in fragments, a problem that may depend on precise settings, although we found computation time and memory use increased significantly at such attempts. This is likely caused by triplexator design to find any combinations of triplex-forming sequences, not only local patterns leading to H-DNA.

*To whom correspondence should be addressed.

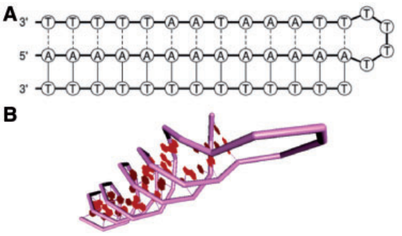

A
3'
T T T T T A A T A A A T T T T T T
5'
A A A A A A A A A A A A A
3'
T T T T T T T T T T T T T
B

Fig. 1. (**A**) 2D diagram and (**B**) 3D model of one of the best scored triplex

2 THE SOFTWARE

The R triplex package is essentially an R interface to the underlying C implementation of a dynamic-programming search strategy of the same name (Lexa *et al*., 2011). The main functionality of the original program was to detect the positions of subsequences in a much larger sequence capable of folding into an intramolecular triplex (H-DNA) made of as many canonical nucleotide triplets as possible. We extended this basic functionality to include the calculation of exact base pairing in the triple helices. This allowed us to include visualization, showing the exact base pairing in 1D, 2D or 3D (see Section 3). The created package takes advantage of the existing Bioconductor infrastructure. For example, the triplex search method uses the *DNAString* object as input. As a result, all available genomes (*BSgenomes* objects) can be easily analyzed. As for the output, identified triplexes are stored in data objects of a class based on *XStringViews*. Thus, all other libraries or methods working with *IRanges* can be applied to triplexes as well. Alternatively, the results can be transformed into *GRanges* objects that enable further possibilities, such as visualization of genome tracks using *GenomeGraphs* or export of results to the GFF3 annotation format.

3 USAGE EXAMPLE

In the following example, we load a genomic sequence from one of the *BSGenome* packages, identify potential triplexes with length over eight triplets of nucleotides and score ≥17, create two different visualizations of the best-scored triplex. Finally, we export the identified positions into a genome annotation track (via a GFF3 file) and store the sequences in a FASTA file.

I) Load necessary libraries and genomes.

```
> library(triplex)
> library(BSgenome.Celegans.UCSC.ce10)
```

II) Search for potential triplex positions and display the results.

```
> t <- triplex.search(Celegans[["chrX"]],
+                     min_score=17,min_len=8)
> t
  Triplex views on a 17718866-letter DNAString subject
subject: CTAAGCCTAAGCCTAAGCCTAAGCCTAAGCCTAA...TAGGCTTAGGCTTAGGCTTAGGCTTAGGCTTAGG
triplexes:
          start width score  pvalue ins type s
    [1]     762    28    17 6.5e-04   0    4 - [TCTAAAAGACACACAATTTAGAAAAAAA]
    [2]    1160    26    17 3.7e-04   0    7 + [ACAAAAACTTCATCAACAAGAAAAAA]
    ...
[20033] 17715172    29    17 3.7e-04   0    6 + [AAAAAAAAGTGAAAAAAACTGAATTTCAT]
[20034] 17718247    27    17 3.7e-04   0    6 + [AAAAAAAAACACTTAAACATAAAACTA]
```

III) Sort the results by score and display the best-scoring non-trivial triplex. Graphical output is shown in Figure 1.

```
> ts <- t[order(score(t),decreasing=TRUE)]
> triplex.diagram(ts[1])
> triplex.3D(ts[1])
```

IV) Export the results as GFF3 and FASTA files.

```
> library(rtracklayer)
> export(as(t, "GRanges")," test.gff", version="3")
> writeXStringSet(as(t, "DNAStringSet"), file="test.fa",
+                 format="fasta")
```

4 CONCLUSION

We present a new R/Bioconductor package that integrates our previously defined algorithm for identification of triplex-forming sequences with two new methods of their visualization (2D diagram and 3D model). The created package uses existing Bioconductor infrastructure in such way that available genomes (*BSGenomes*) can easily be used as input. The identified triplexes can be further analyzed as *IRanges* or *GRanges* objects (and optionally exported into GFF3 or FASTA file). In connection with R language and existing libraries for statistical analysis, the package represents powerful tool for molecular biologists interested in analysis of non-canonical DNA structures such as triplexes.

Funding: Framework of IT4Innovations project (CZ.1.05/1.1.00/02.0070) funded by the EU Operational Programme 'Research and Development for Innovations', MSMT Grants (No.0021630528) 'Security-Oriented Research in Information Technology', and (LA09016) 'Participation of CR in ERCIM', and BUT grant (FIT-S-11-1) 'Advanced secured, reliable and adaptive IT'.

Conflict of Interest: none declared.

REFERENCES

Buske,F.A. *et al.* (2012) Triplexator: detecting nucleic acid triple helices in genomic and transcriptomic. *Genome Res*., **22**, 1372–1381.

Cer,R. *et al.* (2012) *Searching for Non-B DNA-Forming Motifs Using nBMST (Non-B DNA Motif Search Tool)*. Curr. Protoc. Hum. Genet., Chapter 18. Unit 18.7.1–22.

Hoyne,P.R. *et al.* (2000) Searching genomes for sequences with the potential to form intrastrand triple helices. *J. Mol. Biol*., **302**, 797–809.

Htun,H. and Dahlberg,J.E. (1989) Topology and formation of triple-stranded H-DNA. *Science*, **243**, 1571–1576.

Lexa,M. *et al.* (2011) A dynamic programming algorithm for identification of triplex-forming sequences. *Bioinformatics*, **27**, 2510–2517.

Rooney,S.M. and Moore,P.D. (1995) Antiparallel, intramolecular triplex DNA stimulates homologous recombination in human cells. *Proc. Natl Acad. Sci. USA*, **92**, 2141–2144.

Soyfer,V.N. and Potaman,V.N. (1995) *Triple-Helical Nucleic Acids*. Springer-Verlag, New York.

Walter,A. *et al.* (2001) Evidence for a DNA triplex in a recombination-like motif: I. recognition of Watson-Crick base pairs by natural bases in a high-stability triplex. *J. Mol. Recognit*., **14**, 122–139.

A.3 Paper III

pqsfinder: an exhaustive and imperfectiontolerant search tool for potential quadruplexforming sequences in R

Bioinformatics, 33(21), 2017, 3373–3379
doi: 10.1093/bioinformatics/btx413
Advance Access Publication Date: 4 July 2017
Original Paper

Sequence analysis

pqsfinder: an exhaustive and imperfection-tolerant search tool for potential quadruplex-forming sequences in R

Jiří Hon[1], Tomáš Martínek[1], Jaroslav Zendulka[1] and Matej Lexa[2,]*

[1]IT4Innovations Centre of Excellence, Faculty of Information Technology, Brno University of Technology, 61266 Brno, Czech Republic and [2]Department of Information Technology, Faculty of Informatics, Masaryk University, 60200 Brno, Czech Republic

*To whom correspondence should be addressed.

Associate Editor: John Hancock

Received on April 20, 2017; revised on June 6, 2017; editorial decision on June 19, 2017; accepted on June 23, 2017

Abstract

Motivation: G-quadruplexes (G4s) are one of the non-B DNA structures easily observed *in vitro* and assumed to form *in vivo*. The latest experiments with G4-specific antibodies and G4-unwinding helicase mutants confirm this conjecture. These four-stranded structures have also been shown to influence a range of molecular processes in cells. As G4s are intensively studied, it is often desirable to screen DNA sequences and pinpoint the precise locations where they might form.

Results: We describe and have tested a newly developed Bioconductor package for identifying potential quadruplex-forming sequences (PQS). The package is easy-to-use, flexible and customizable. It allows for sequence searches that accommodate possible divergences from the optimal G4 base composition. A novel aspect of our research was the creation and training (parametrization) of an advanced scoring model which resulted in increased precision compared to similar tools. We demonstrate that the algorithm behind the searches has a 96% accuracy on 392 currently known and experimentally observed G4 structures. We also carried out searches against the recent G4-seq data to verify how well we can identify the structures detected by that technology. The correlation with *pqsfinder* predictions was 0.622, higher than the correlation 0.491 obtained with the second best G4Hunter.

Availability and implementation: http://bioconductor.org/packages/pqsfinder/ This paper is based on pqsfinder-1.4.1.

Contact: lexa@fi.muni.cz

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

DNA sequences capable of forming alternative secondary structures, called non-B DNA, have long been at the center of research interest because of their possible biological functions (Du *et al.*, 2013) and their involvement in mutagenesis and disease (Bacolla and Wells, 2009). Instead of forming canonical B-DNA helices with Watson-Crick base pairing, these regions of DNA can engage in different types of base pairing and form cruciforms, triplexes (or H-DNA), G-quadruplexes (G4s), i-motifs and a few other alternative structures (Wells, 2007). After previous work on algorithms and practical solutions to identify triplex DNA (Hon *et al.*, 2013; Lexa *et al.*, 2011), we focus here on identifying potential quadruplex-forming sequences (PQS).

As evidenced by sequencing (Chambers *et al.*, 2015), as well as a large number of other experimental and *in silico* studies, PQS are found in high numbers in eukaryotic genomes (Huppert, 2005; Lexa et al., 2014). They are implicated in several genome-wide processes, mostly as positive or negative regulators of transcription (Rhodes and Lipps, 2015), negative regulators of replication which require specialized helicases for the processes to continue (Mendoza et al., 2016) and may be dispersed into critical locations of the genome by the activity of transposable elements (Kejnovsky and Lexa, 2014).

Today, several software tools for identification of PQS in biological sequences are available. The oldest and most commonly used algorithms are based on a simple folding rule representing four runs of guanines separated by relatively short loops (or spacers). These include quadparser (Huppert, 2005), QGRS Mapper (D'Antonio and Bagga, 2004; Kikin et al., 2006) and Quadfinder (Scaria et al., 2006). The folding rule used in these tools is usually of the form G{3,6}.{1,8}G{3,6}.{1,8}G{3,6}.{1,8}G{3,6} reflecting the fact that PQS with short loops and four perfect G runs form the most stable G4s *in vitro*. These tools consider only sequences that match the sequence formula perfectly.

In recent years, different *in vitro* experiments have confirmed the existence of imperfect G4s (Mukundan and Phan, 2013). They have also been explored *in silico* by molecular dynamics (Varizhuk et al., 2017). As a result, new tools for prediction of imperfect G4s began to be developed. Such tools include TetraplexFinder/QuadBase2 (Dhapola and Chowdhury, 2016), ImGQfinder (Varizhuk et al., 2014) and G4Hunter (Bedrat et al., 2016). For example, TetraplexFinder considers potential bulges of defined length in runs of three guanines, while ImGQfinder considers the possibility of a single bulge or mismatch in a wider variety of guanine run lengths. Finally, G4Hunter does not define individual defect types, but uses a simple encoding and statistics over a sliding window, that can accomodate different types of defects.

It has also been discovered that a given DNA segment (sequence) can form several overlapping G4s, by definition mutually exclusive, where individual nucleotides in the sequence compete with each other for binding via Hoogsteen bonds (Agrawal et al., 2014). In these cases, it is very useful to have a tool for predicting all overlapping instances and evaluate them with scores that correlate with the propensity for G4 formation. The only tool predicting overlapping G4s and at the same time capable of assigning scores to their individual instances is QGRS Mapper. Its score function considers the number of Gs in each run, loop lengths as well as the difference in loop lengths. Features of existing software tools for PQS identification are summarized in Table 1.

In this paper, we introduce an R package and the underlying algorithm for PQS detection that addresses certain shortcomings of the available tools.

Five main ideas projected into the package functioning are to: (i) allow imperfections in PQS as mismatches or bulges in G runs and excessively long loops between the G runs, (ii) provide a PQS score that is closely related to G4 stability, (iii) give the user a choice between reporting all overlapping PQS and/or only the locally best, (iv) provide the overall number (density) of possible PQS conformations covering each position in the input sequence and (v) allow users to define their own criteria for matching and scoring, overriding the defaults determined by calculations in this paper.

The package and the algorithm were called *pqsfinder* and accepted into Bioconductor (Huber et al., 2015) in April 2016. Here, we explain how the ideas were implemented in the package and apart from tuning its default parameters and settings, we show how *pqsfinder* predictions relate to recently carried out G4 sequencing (also called G4-seq or G-seq) (Chambers et al., 2015).

2 Approach and algorithm

The main principle of the algorithmic approach presented here is based on the fact that monomolecular G4 structures arise from compact sequence motifs composed of four consecutive and possibly imperfect guanine runs (G runs) interrupted by loops of semi-arbitrary lengths.

The algorithm first identifies four consecutive G run sequences (G run quartet). Subsequently, it examines the potential of such G run quartet to form a stable G4 and reports a corresponding quantitative score.

The *pqsfinder* algorithm can be divided into three logical steps: (i) identification of all possible G run quartets, (ii) score assignment and (iii) overlap resolution. All three parts are described in the following sections.

2.1 Identification of all possible G run quartets

The first G run is matched freely in the sequence by a regular expression G{1,10}.{0,9}G{1,10} with limited minimal and maximal length. This regular expression allows us to match imperfect G runs containing both mismatches and bulges while requiring at least two guanines. The remaining three G runs are matched by the same regular expression with the following additional constraints: (i) each subsequent G run must lie beyond the 3'-end of the previous one (no overlap), (ii) the distance of each G run to the previous G run must be in the range of minimal and maximal loop length and at most one loop is allowed to have zero length (Marusic et al., 2013) and (iii) each G run has to fit in a sequence window defined by the first G run starting position and the user-defined maximal PQS length. These constraints are summarized in Figure 1.

As regular expressions are able to capture only one match (usually the maximal one), to list all possible combinations we use a backtracking approach. After four initial G runs are matched and processed, the last successfully matched G run is shortened by one

Table 1. Feature comparison of existing tools for PQS identification

Name	Model	Overlaps	Imperf.	Score	Avail.
quadparser	Folding rule	✓	✗	✗	✗
QGRS Mapper	Folding rule	✓	✗	✓	Web
Quadfinder	Folding rule	✓	✗	✗	✗
ImGQfinder	Folding rule	✓	✓[a]	✗	Web
TetraplexFinder	Regular expression	✓	✓[b]	✗	Web
G4Hunter	Sliding window	✗[c]	✓	✓	R script

[a] ImGQfinder allows at most one imperfection.

[b] TetraplexFinder supports only bulges of fixed length between 0 and 7.

[c] G4Hunter model inherently merges overlapping and neighbouring PQS. For this reason, the boundaries of individual PQS are not well-defined.

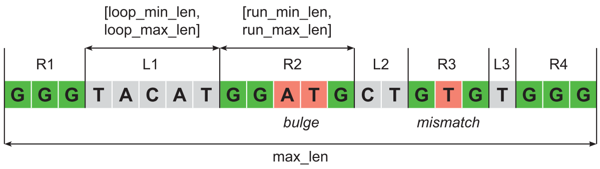


Fig. 1. PQS constraints. Every PQS consists of two types of elements: G runs (R1–4) and loops (L1–3). The minimal and maximal length of each element type is constrained by the corresponding options depicted in the picture as well as the overall PQS length. All these options can be freely customized when using the *pqsfinder* package

nucleic acid base from the end and if it is still a valid G run, the algorithm proceeds normally to scoring and overlap resolution. On the other hand, if the shortened G run is not valid, the algorithm tracks back to the previous successfully matched G run and applies the same shortening modification. In this case, if the modified G run is valid, the algorithm proceeds to match all the following G runs again. Once the backtracking procedure gets to the first G run and finds its shortened variant to be invalid, the whole process of G run identification is rerun from position one after the starting position of the first G run. The backtracking procedure increases the computational complexity of the search, but allows us to rigorously model the competition between overlapping PQS.

2.2 Score assignment

The *pqsfinder* scoring scheme was designed to quantitatively approximate the relationship between G4 sequence and the stability of its structure. While the scoring function is purely empirical, we intentionally chose an approach where the score is modular and, obtained by addition of scores representing the binding affinities of smaller regions within the G4. This kind of approach has already been proven to work for simpler DNA structures, such as nucleic acid duplexes and hairpins. (SantaLucia, 2012; Zuker, 2003)

The first part of the scoring scheme quantifies the quality of individual G runs. It awards the PQS a score for each G-tetrad stacking and penalizes mismatches and bulges in G runs.

The scoring is then defined by Equation 1, where N_t is the number of tetrads, B_t is a G-tetrad stacking bonus, N_m is the number of inner mismatches, P_m is mismatch penalization, N_b is the number of bulges, P_b is bulge penalization, F_b is bulge length penalization factor, L_{bi} is the length of the i-th bulge and E_b is bulge length exponent.

$$S_r = (N_t - 1)B_t - N_m P_m - \sum_{i=1}^{N_b} P_b + F_b L_{bi}^{E_b} \qquad (1)$$

However, discrimination between bulges and mismatches can be a demanding task requiring multiple sequence alignment. To avoid this, we made two simplifying assumptions that allowed us to efficiently analyze bulges and mismatches by only counting lengths of G runs and their G content. First, we require at least one G run to be perfect (consisting of just guanines). Second, we limit the number of imperfections to one per G run. Based on the available literature, we consider bulges and long loops to be strong destabilizers of G4s and do not expect more than a few of these imperfections to be possible at the same time.

In the scoring procedure, a perfect G run is taken as a reference and other G runs are assessed relatively to the reference. A G run is classified as mismatched, if it has the same length as the reference and the G content lower by one. When a G run has a greater length than the reference and at least the same G content, it is classified as bulged. Finally, all G runs can only be either perfect, mismatched or bulged. Other cases are considered to be invalid G runs. When there are multiple perfect G runs present, the shortest one is used as the reference.

The second part of the scoring scheme quantifies the destabilizing effect of the loops on G4 stability. At this time we have no mechanistic understanding of possible loop sequence and length effects. Hence, we limit ourselves to an empirical formula that can accommodate some of the observations made by Guédin *et al.* (2010). Loop length mean L_m is multiplied by the factor F_m and raised to the power of E_m. Complete scoring function is then expressed by Equation 2.

$$S = \max(S_r - F_m L_m^{E_m}, 0) \qquad (2)$$

F_m and E_m are numerical parameters that empirically model the relationship between loop lengths and their destabilization effects on the quadruplex. These permit a non-linear relationship, while their values are derived by fitting the model to experimental results (see Section 4). S_r is the value from Equation 1.

2.3 Overlap resolution

The overlap resolution is an iterative process that is designed to always prefer dominant PQS. First, all PQS sharing the highest obtained score are selected (in subsequent iterations, PQS sharing the highest remaining score are used). Second, the selected PQS are processed one by one in the order of their increasing starting position as follows: (i) if the current PQS overlaps the previous PQS, the current PQS is removed, (ii) if the current PQS is completely included in the previous PQS, the previous PQS is removed. Third, all lower-scoring PQS overlapping with any of the remaining selected PQS are discarded. Fourth, all selected PQS are reported and removed. Fifth, the next iteration begins again with the remaining PQS. Iterations continue until all PQS are checked (either reported or removed).

We implemented the process above effectively in order to reduce the memory usage. The main optimization idea is to run the iterative process progressively as the identification algorithm proceeds through the sequence. As a result, only a small set of recently identified overlapping PQS has to be in memory.

3 Implementation

The *pqsfinder* package was created following recommended practices for R/Bioconductor packages and all functions are well-documented within the inline R documentation system. A detailed user guide with convenient examples was also prepared as a package vignette. Source code is written in both R and C++, each having its own important role in the package architecture.

The R code implements the interface that is needed for a seamless user interaction within the Bioconductor framework, relying on the following R packages: Biostrings (Pagès *et al.*, 2016a), GenomicRanges, IRanges (Lawrence *et al.*, 2013), S4Vectors (Pagès *et al.*, 2016b), Rcpp (Eddelbuettel and François, 2011) and BH (Eddelbuettel *et al.*, 2016). The package provides one main function *pqsfinder* for running the PQS search algorithm and several secondary functions that operate on the search results.

The central data structure for results is the *PQSViews* class which is derived from the *XStringViews* class from the Biostrings package. It maintains the sequence coordinates of the identified PQS along with other useful metadata: (i) score, (ii) strand, (iii) number of tetrads, (iv) number of bulges, (v) number of mismatches and (vi) loop lengths.

Table 2. Overview of *pqsfinder* options

Group	Name	Description
Filters	*strand*	Strand symbol: +, – or * (both).
	overlapping	Enables overlapping PQS.
	max_len	Maximal PQS length.
	min_score	Minimal PQS score.
	run_min_len	Minimal G run length.
	run_max_len	Maximal G run length.
	loop_min_len	Minimal loop length.
	loop_max_len	Maximal loop length.
	max_bulges	Maximal number of bulges.
	max_mismatches	Maximal number of mismatches.
	max_defects	Maximal number of all defects.
Scoring	*tetrad_bonus*	G-tetrad stacking bonus B_t.
	mismatch_penalty	Inner mismatch penalization P_m.
	bulge_penalty	Bulge penalization P_b.
	bulge_len_factor	Bulge length penal. factor F_b.
	bulge_len_exponent	Bulge length penal. exponent E_b.
	loop_mean_factor	Loop mean penal. factor F_m.
	loop_mean_exponent	Loop mean penal. exponent E_m.
Advanced	*run_re*	G run regular expression.
	custom_scoring_fn	User-defined scoring function.
	use_default_scoring	Enables internal scoring system.
	verbose	Enables detailed text output.

This aside, the *PQSViews* object provides access to two additional vectors. The first is a *density* vector—for each sequence position it gives the number of different PQS conformations overlapping that position. The second vector *maxScores* reports the PQS quality along the sequence—for each sequence position it gives the maximal score of all PQS overlapping that position. We consider these two vectors particularly useful as additional information to the exact PQS coordinates and metadata. The *density* and *maxScores* vectors can be easily used to discriminate low-complexity regions (full of guanines) that inherently allow a large amount of folded PQS conformations from regions that on the other hand contain a singular high-scoring PQS.

The main PQS search logic is implemented purely in the C++ language for speed since the algorithm is based on an exhaustive search of the PQS topological space and it is computationally intensive by definition. The Rcpp library was used to easily link the C++ code with R scripts. We also employed the Boost regular expression library (Maddock, 2016) to match individual G runs. However, we soon realized that the general regular expression engine has a significant overhead and is too slow for our needs. For this reason, we implemented an optimized matching function for the default G run regular expression. At the same time, we are linking the Boost library for the case where users would like to use their own definition of a G run using an alternative regular expression.

3.1 Customization

Since we strongly support the Bioconductor goal to further scientific understanding by producing extensible, scalable and interoperable software, we designed *pqsfinder* to be easily customizable. The users can tweak the algorithm options for their personal needs or test new hypotheses about PQS conformations and develop novel innovative scoring schemes. Supported options are divided into three logical groups: (i) filters, (ii) scoring and (iii) advanced (see Table 2).

Filter options control the main algorithmic constraints (see Fig. 1). These have great impact on the algorithm sensitivity and speed. All PQS that do not satisfy the basic constraints are excluded immediately and do not proceed further to the scoring step.

Scoring options include all the constants that appear in the scoring Equations 1 and 2. By default, these constants are set to reasonable values as described in the next section and its modification is recommended only to users who would like to bias the scoring systems towards a specific type of G4 or to refine the constants on novel data.

Advanced options allow to get full control over the search algorithm by providing alternative G run regular expression and scoring function. However, the custom scoring function can negatively influence the overall algorithm performance, particularly on long sequences, since there is a significant overhead linked to the calling of custom R function instead of efficient inline C++ implementation. Thus, this feature is recommended only for rapid prototyping of novel scoring techniques, which can be later implemented efficiently in C++ and delivered in the next version of the *pqsfinder* package.

4 Model training

As described in the foregoing section, the scoring model requires several constants to be chosen (see *scoring* group in Table 2). It is, however, very difficult to estimate these parameters. For this reason, we decided to construct a training set from available experimental data and search for a setting that gives the best performance on these data. The dataset construction process and parameter-search algorithm are discussed in detail in this section.

4.1 Existing datasets

Methods for G4 prediction are usually evaluated on a set of experimentally verified (*in vitro*) G4s, extracted from different publications. For example, a recently published method G4Hunter involved collecting a set of 392 experimentally verified G4s consisting of 298 positive and 94 negative samples (later referred to as Lit392).

However, these datasets have several disadvantages: (i) they are unbalanced regarding the number of positive and negative samples, (ii) significant number of items differ only by a single mutation and (iii) datasets are very small and cover only a small proportion of possible G4 conformations given all the possible loop lengths, bulges, mismatches and other defects.

On the other hand, (Chambers *et al.*, 2015) recently published a novel approach for high throughput sequencing of DNA G4 structures called G4-seq. The technique detects noisy sequences that emerge on treatment of DNA samples with K^+ or PDS (pyridostatin, a chemical G4 stabilizer). As a result of this technology, the authors released a track (in BED format) that shows the propensity of reference Human DNA sequence (*hg19*) to form G4s.

This track has two disadvantages. First, it only shows the level of mismatches at given sequence positions that were observed during the sequencing process. Hence, in reality, we have no evidence that a G4 has been formed, but based on the G4-seq method the level of mismatches should show high correlation with the probability that the sequence forms the G4 structure. Second, as the G4 structure is formed during sequencing, the level of mismatches remains high, until the end of the sequenced read, even downstream of the actual G4 structure. As a result, the BED file constructed by mapping the reads onto the reference sequence, can be affected by this ‘memory effect’.

Despite these disadvantages, the G4-seq dataset is extremely valuable, because it shows the G4 structure propensity for the entire human genome and thus it covers many more possible conformations and imperfect structures (including long loops and bulges) than any dataset extracted from the published literature.

Based on these facts we decided to use a subset of G4-seq data for training of the *pqsfinder* scoring model. We then used two additional independent datasets for testing: Lit392 and a different (non-overlapping with training data) subset of G4-seq data. The whole process was operated as follows:

1. We prepared independent training and test sets from G4-seq data.
2. We trained *pqsfinder* parameters on G4-seq training set.
3. We selected those parameters that performed best on the G4-seq training set.
4. Finally, the selected *pqsfinder* parameters were evaluated and compared to other tools on the Lit392 dataset and G4-seq test set.

In the following subsection, individual steps of this procedure are described in more detail.

4.2 Preparation of the training and test sets

From the G4-seq data, we used BED files representing the level of mismatches from two experimental treatments. In the first treatment, the authors stabilized G4s using K^+ while in the second case they used PDS. In both cases, measurements were done on both DNA strands separately resulting in four BED files (two treatments with two strands each).

In the first step, as the K^+ and PDS measurements do not cover 100% of *hg19* genome, we identified only those DNA fragments where both K^+ and PDS measurements were available. Then, we filtered out fragments shorter than 10 kbp and longer fragments were trimmed to 10 kbp. In the next step, we combined K^+ and PDS BED files by calculating the average value from both treatments. Subsequently, we filtered out those fragments that did not include a significant level of mismatches (where the averaged level of mismatches from K^+ and PDS never exceeded threshold 40). In order to eliminate cases where potential G4 overlapped the beginning or the end of the fragment, we also filtered out those fragments that included a significant level of mismatches in the first 30 bp or the last 30 bp of the fragment.

The described procedure was applied to each strand separately. Finally, 1100 fragments were chosen at random, 100 as G4-seq training set (for a total of 1 Mbp) and 1000 as G4-seq test set (for a total of 10 Mbp). Both datasets are available as Supplementary Data.

4.3 Training of scoring parameters on the G4-seq training set

We used the genetic algorithm implemented in the R package GA (Scrucca, 2013) as a method for parameter-space exploration and training. In order to make the exploration process easier, the G-tetrad stacking bonus was fixed at 40. The remaining scoring options were trained. Their names, number of bits allocated in GA chromosome and ranges of values considered are summarized in Table 3. Total GA chromosome length was 33 bits. Other *pqsfinder* options were fixed to the default values.

To evaluate fitness, we calculated Pearson's correlation coefficient between the vector *maxScores* generated by *pqsfinder* (see section 3) and the averaged level of mismatches from K^+ and PDS treatments of G4-seq training set. More specifically, maximal values of the *pqsfinder* score were calculated for all positions of all DNA fragments in the training set and these values were correlated with appropriate positions in the G4-seq training set (experimentally verified level of mismatches). The basic idea behind this fitness function is: the higher the correlation coefficient between *pqsfinder* score and G4-seq mismatch level, the better the prediction of putative G4 structures will be.

Table 3. Trained parameters and their encoding in chromosome

Name	Bits	Range	Step	Result
bulge_penalty	6	0–63	1	20
mismatch_penalty	6	0–63	1	28
bulge_len_factor	5	0–3.1	0.1	0.2
bulge_len_exponent	5	0–3.1	0.1	1
loop_mean_factor	6	3–9.3	0.1	6.6
loop_mean_exponent	5	0–3.1	0.1	0.8

A genetic algorithm was set up with the following parameters: (i) population size 24, (ii) probability of crossover 0.5, (iii) probability of mutation 0.5 and (iv) number of generations 200. During the exploration process, we used monitor function and recorded 1157 unique combinations of parameters and their fitness values.

As the final parameters, we selected the combination with the maximal fitness value. Concrete values of selected parameters are listed in Table 3 (column *Result*). The table of all explored parameter combinations and their fitness values is available as Supplementary Data.

5 Results

In the first step, we compared *pqsfinder* to other tools capable to predict whether a given sequence can form a G4 or not. As candidate tools that are still working and available online/offline, we selected: G4Hunter, QGRS Mapper, TetraplexFinder and ImGQfinder. We applied these to a recently published dataset (Bedrat *et al.*, 2016) containing 392 *in vitro* verified G4s (Lit392), originally used to test G4Hunter.

In the next step, we configured and executed the selected tools with the following parameters. (i) *pqsfinder* was executed with the parameters that had the best fitness value on the G4-seq training set. (ii) G4Hunter was executed with the default parameters. (iii) QGRS Mapper was executed with the most relaxed parameters, i.e. minimal G run length was 2, loop length was in the range 0 to 36 and maximal length was 45. As *pqsfinder*, G4Hunter and QGRS Mapper report scores, to calculate accuracy and Matthews correlation coefficient (MCC), we always systematically found a threshold that resulted in the highest possible values for each tool. Interestingly, we found out that for G4Hunter, the threshold 0.71 works even better than thresholds 1.0, 1.2 and 1.5 that are recommended by the authors. (iv) TetraplexFinder was executed with the following combinations of parameters: G run length 2 and 3, greedy and non-greedy approach, bulge length in the range 0 to 7 and maximal loop length 50. Of all possible TetraplexFinder parameter combinations, only the best ones are reported in Table 4. (v) ImGQfinder was executed with G run length in the range 2 to 5, maximal loop length 25 and number of defects 0 and 1. Again, only the best combinations are presented in Table 4.

Finally, for all selected tools and their configurations, we measured basic performance characteristics, namely accuracy (ACC) and Matthews correlation coefficient (MCC). For tools that report a score or allow us to specify a threshold, we also measured the area under the ROC curve (AUC). The results are summarized in Table 4. Since the Lit392 dataset is unbalanced, MCC is the most relevant value. As we can see, *pqsfinder* outperformed other tools significantly.

Table 4. Performance comparison of different tools on Lit392 dataset

Tool	Configuration	ACC	MCC	AUC
pqsfinder	Best on G4-seq training set	0.964	0.902	0.975
G4Hunter	Default	0.952	0.865	0.969
QGRS Mapper	g≥2, *ll* = 36, *l* = 45	0.954	0.872	0.968
TetraplexFinder	*g* = 2, *ll* = 50, *gr*, *bl* = 0	0.946	0.850	–
TetraplexFinder	*g* = 2, *ll* = 50, *ngr*, *bl* = 0	0.946	0.850	–
ImGQfinder	*g* = 2, *ll* = 25, *d* = 0	0.941	0.835	–
ImGQfinder	*g* = 2, *ll* = 25, *d* = 1	0.918	0.767	–

Note: The meaning of the configuration options is as follows: *t* is threshold, *g* is G run length, *ll* is maximal loop length, *l* is maximal G4 length, *gr* is greedy approach, *ngr* is non-greedy approach, *bl* is bulge length and *d* is number of defects. For tools that report score (*pqsfinder*, G4Hunter and QGRS Mapper), we systematically determined thresholds that resulted in the highest possible ACC and MCC. We also found that for tools without scoring system (TetraplexFinder and ImGQfinder) it is always better to disable imperfections.

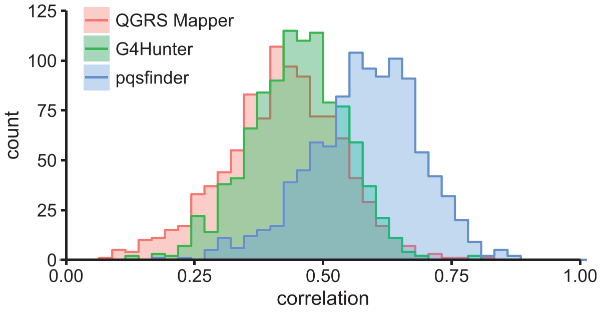


Fig. 2. Histogram of correlation coefficients for QGRS Mapper, G4Hunter and *pqsfinder* on the G4-seq test set fragments. The correlation was measured between the averaged level of mismatches of the G4-seq test set fragments (see Section 4.2) and the vector of maximal scores predicted by each tool. While histograms of QGRS Mapper and G4Hunter correlations are almost the same, the histogram of *pqsfinder* correlations is much more positively skewed

Subsequently, we identified tools capable of predicting overlapping G4s and assigning them a score. Only those tools could also be evaluated on the G4-seq test set. The basic idea behind this test is to calculate all possible overlapping G4s for a given sequence and extract the characteristics of maximal score values (for every sequence position maximal score of all overlapping G4s is selected). Such characteristic can then be correlated with the level of mismatches at the same positions of the G4-seq test set. From the set of available tools, only the QGRS Mapper and G4Hunter met the requirement. As a dataset, we used G4-seq test set consisting of 1000 randomly selected DNA fragments with length of 10 kbp (procedure for dataset construction is described in Section 4.2).

In the next step we configured and executed selected tools with the following parameters: *pqsfinder* was executed with parameters trained on G4-seq training set. G4Hunter was executed with all thresholds between 0 and 4 (with step 0.05). Predicted G4s were refined and merged together. QGRS Mapper was evaluated with the most relaxed parameters as before, i.e. minimal G run length is 2, loop length is in range 0 to 36 and maximal length is 45. For results from each tool, the characteristic of maximal score value was calculated and compared with the G4-seq test set. This comparison was done in two ways. First, Pearson correlation coefficient was calculated for every fragment separately. As the result, we got a distribution of correlation coefficients with individual means and standard deviations (see Fig. 2 and Table 5, columns *CC mean* and *CC SD*). Second, Pearson correlation coefficient was calculated for all fragments joined together to get a single overall value (see Table 5, column *Overall CC*). As we can see, the *pqsfinder* significantly outperformed other tools.

Table 5. Comparison of correlation coefficient (CC) statistics for different tools

Tool	CC mean	CC SD	Overall CC
pqsfinder	0.583	0.106	0.622
G4Hunter	0.450	0.093	0.491
QGRS Mapper	0.422	0.112	0.479

Note: The individual CCs were measured between the averaged level of mismatches of the G4-seq test set fragments (see Section 4.2) and the vector of maximal scores predicted by each tool. Overall CC was calculated between concatenated averaged level of mismatches of all G4-seq test fragments and concatenated vector of the corresponding predicted maximal scores.

6 Discussion

The objective of the tools for G4 prediction is to model the complex relationship between DNA sequence and G4 structure. Despite our ability to model this relationship directly at the molecular level, using for example molecular dynamics Amber tool (Salomon-Ferrer *et al.*, 2013), this approach is computationally demanding and the accuracy of the state-of-the-art force fields is still limited. For these reasons, existing tools for G4 prediction use much simpler models.

The majority of tools are based on a simple folding rule and are very fast, but do not allow for possible defects (mismatches and bulges) easily. There are tools, such as TetraplexFinder and ImGQfinder that allow for imperfections in G-quadruplexes. However, without a properly trained scoring model this can easily lead to a large number of false positives. These tools performed better in our tests when imperfections were limited or not allowed at all.

A very interesting approach allowing imperfections that is based on specific encoding and simple statistic over a sliding window was implemented in G4Hunter. Despite its simplicity, it shows very good performance characteristics. Unfortunately, we believe that such simple encoding and statistics cannot reveal all complex relationships between sequence and G4 stability, and thus the accuracy of such approach is limited.

On the other hand, the approach proposed in this article that combines pattern matching and detailed inspection of possible defects is configurable and easily extensible. Using advanced options, it can be quickly customized to detect novel and experimental G4 types that are currently not commonly studied or might be discovered in the future. One such example is the recently postulated interstrand G4s (Kudlicki, 2016) or G4s formed in *cis*, as proposed by Hegyi (2015).

By default, the *pqsfinder* provides a scoring function that was trained on G4-seq experimental data and performs better than competing tools. We are aware that G4-seq data essentially represent conditions *in vitro* and may not necessarily be directly related to the ability of G4s to form *in vivo*, but our current view is that *in vivo* G4 formation is a function of their *in vitro* stability. Therefore, G-seq experimental data is the best publicly available dataset we could find at this moment.

However, detailed inspection and modularity are at the cost of lower processing speed. In the extremely sensitive configuration having the minimal G run length set to 2, the algorithm is able to process approximately 4 kb per second on current hardware. For example, *pqsfinder* running time on the G4-seq test set (in total 10 Mbp) was around 40 minutes. When the minimal G run length is increased by one, the speed is usually more than doubled. We do not consider the speed limitations to be critical. For the most frequently studied sequences, *pqsfinder* results can be precomputed and provided to many users, for example as an R data file or a GFF3-formatted file.

7 Conclusion

We created a PQS detection tool with a sequence scoring function that has a moderate number of tunable parameters reflecting sequence properties previously associated with observed G4s or their destabilization (number of Gs in G runs, loop length, presence of mismatches and bulges). To model G-quadruplexes and search for the responsible sequences, we selected a mix of known and novel approaches that give the *pqsfinder* several desirable characteristics. In our tests it achieved the best accuracy on both experimentally verified G-quadruplexes (Lit392) and the independent part of G4-seq data (none of these datasets were used for training). The *pqsfinder* estimates the total number of possible local conformations, accounts for competition between them and allows for imperfections with a sound, carefully trained structure-based scoring model. The presented model was trained on a subset of G4-seq data that represents the largest set of experimentally verified quadruplex-forming sequences available so far and includes a wide variety of imperfections. This new tool also evaluates all the competing conformations and can be easily expanded or modified for newly discovered rules and scoring functions in future. We provide evidence that the *pqsfinder* is a convenient R/Bioconductor package compatible with many other packages available in this environment.

Acknowledgements

Access to computing and storage facilities owned by parties and projects contributing to the National Grid Infrastructure MetaCentrum provided under the programme ‘Projects of Large Research, Development, and Innovations Infrastructures’ (CESNET LM2015042) is greatly appreciated.

Funding

This work was supported by the Czech Science Foundation [15-02891S to M.L.], IT4Innovations excellence in science project of the Ministry of Education, Youth and Sports of the Czech Republic [IT4I XS – LQ1602 to T.M. and J.Z.] and ICT tools, methods and technologies for smart cities project of the Brno University of Technology [FIT-S-17-3964 to J.H.].

Conflict of Interest: none declared.

References

Agrawal,P. *et al.* (2014) The major G-quadruplex formed in the human BCL-2 proximal promoter adopts a parallel structure with a 13-nt loop in K^+ solution. *J. Am. Chem. Soc.*, **136**, 1750–1753.

Bacolla,A. and Wells,R.D. (2009) Non-B DNA conformations as determinants of mutagenesis and human disease. *Mol. Carcinogenesis*, **48**, 273–285.

Bedrat,A. *et al.* (2016) Re-evaluation of G-quadruplex propensity with G4Hunter. *Nucleic Acids Res.*, **44**, 1746–1759.

Chambers,V.S. *et al.* (2015) High-throughput sequencing of DNA G-quadruplex structures in the human genome. *Nat. Biotechnol.*, **33**, 877–881.

D’Antonio,L. and Bagga,P. (2004) Computational methods for predicting intramolecular G-quadruplexes in nucleotide sequences. In: *Proceedings of the 2004 IEEE Computational Systems Bioinformatics Conference*, 2004, *CSB 2004*, Institute of Electrical and Electronics Engineers (IEEE). pp. 590–591.

Dhapola,P. and Chowdhury,S. (2016) QuadBase2: web server for multiplexed guanine quadruplex mining and visualization. *Nucleic Acids Res.*, **44**, W277–W283.

Du,X. *et al.* (2013) The genome-wide distribution of non-B DNA motifs is shaped by operon structure and suggests the transcriptional importance of non-B DNA structures in *Escherichia coli*. *Nucleic Acids Res.*, **41**, 5965–5977.

Eddelbuettel,D. and François,R. (2011) Rcpp: Seamless R and C++ integration. *J. Stat. Softw.*, **40**, 1–18.

Eddelbuettel,D. *et al.* (2016) BH: Boost C++ Header Files. R package version 1.62.0-1.

Guédin,A. *et al.* (2010) How long is too long? Effects of loop size on G-quadruplex stability. *Nucleic Acids Res.*, **38**, 7858–7868.

Hegyi,H. (2015) Enhancer-promoter interaction facilitated by transiently forming G-quadruplexes. *Scientific Rep.*, **5**, 9165.

Hon,J. *et al.* (2013) Triplex: an R/Bioconductor package for identification and visualization of potential intramolecular triplex patterns in DNA sequences. *Bioinformatics*, **29**, 1900–1901.

Huber,W. *et al.* (2015) Orchestrating high-throughput genomic analysis with Bioconductor. *Nat. Methods*, **12**, 115–121.

Huppert,J.L. (2005) Prevalence of quadruplexes in the human genome. *Nucleic Acids Res.*, **33**, 2908–2916.

Kejnovsky,E. and Lexa,M. (2014) Quadruplex-forming DNA sequences spread by retrotransposons may serve as genome regulators. *Mobile Genet. Elements*, **4**, e28084.

Kikin,O. *et al.* (2006) QGRS mapper: a web-based server for predicting G-quadruplexes in nucleotide sequences. *Nucleic Acids Res.*, **34**, W676–W682.

Kudlicki,A.S. (2016) G-quadruplexes involving both strands of genomic DNA are highly abundant and colocalize with functional sites in the human genome. *Plos One*, **11**, e0146174.

Lawrence,M. *et al.* (2013) Software for computing and annotating genomic ranges. *PLoS Comput. Biol.*, **9**, e1003118.

Lexa,M. *et al.* (2011) A dynamic programming algorithm for identification of triplex-forming sequences. *Bioinformatics*, **27**, 2510–2517.

Lexa,M. *et al.* (2014) Guanine quadruplexes are formed by specific regions of human transposable elements. *BMC Genomics*, **15**, 1032.

Maddock,J. (2016). *Boost.Regex 5.1.2*.

Marusic,M. *et al.* (2013) G-rich vegf aptamer with locked and unlocked nucleic acid modifications exhibits a unique g-quadruplex fold. *Nucleic Acids Res.*, **41**, 9524–9536.

Mendoza,O. *et al.* (2016) G-quadruplexes and helicases. *Nucleic Acids Res.*, **44**, 1989–2006.

Mukundan,V.T. and Phan,A.T. (2013) Bulges in G-quadruplexes: Broadening the definition of G-quadruplex-forming sequences. *J. Am. Chem. Soc.*, **135**, 5017–5028.

Pagès,H. *et al.* (2016a) *Biostrings: String objects representing biological sequences, and matching algorithms*. R package version 2.42.1.

Pagès,H. *et al.* (2016b) *S4Vectors: S4 implementation of vectors and lists*. R package version 0.12.1.

Rhodes,D. and Lipps,H.J. (2015) G-quadruplexes and their regulatory roles in biology. *Nucleic Acids Res.*, **43**, 8627–8637.

Salomon-Ferrer,R. *et al.* (2013) An overview of the Amber biomolecular simulation package. *Wiley Interdisc. Rev. Comput. Mol. Sci.*, **3**, 198–210.

SantaLucia,J. Jr., (2012) A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics. *Proc. Natl. Acad. Sci. USA*, **95**, 1460–1465.

Scaria,V. *et al.* (2006) Quadfinder: server for identification and analysis of quadruplex-forming motifs in nucleotide sequences. *Nucleic Acids Res.*, **34**, W683–W685.

Scrucca,L. (2013) GA: a package for genetic algorithms in R. *J. Stat. Softw.*, **53**, 1–37.

Varizhuk,A. *et al.* (2014) An improved search algorithm to find G-quadruplexes in genome sequences. *bioRxiv*.

Varizhuk,A. *et al.* (2017) The expanding repertoire of G4 DNA structures. *Biochimie*, **135**, 54–62.

Wells,R.D. (2007) Non-B DNA conformations, mutagenesis and disease. *Trends Biochem. Sci.*, **32**, 271–278.

Zuker,M. (2003) Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.*, **31**, 3406–3415.

A.4 Paper IV

Uneven distribution of potential triplex sequences in the human genome In silico study using the R/Bioconductor package triplex

Uneven distribution of potential triplex sequences in the human genome
In silico study using the R/Bioconductor package triplex

Matej Lexa[1], Tomáš Martínek[2] and Marie Brázdová[3]
[1]*Faculty of Informatics, Masaryk University, Botanická 68a, 60200 Brno, Czech Republic*
[2]*Faculty of Information Technology, Brno Technical University, Božetěchova 1/2, 61266 Brno, Czech Republic*
[3]*Biophysical Institute of the Czech Academy of Sciences, Královopolská 135, 61265 Brno, Czech Republic*
lexa@fi.muni.cz, martinto@fit.vutbr.cz, maruska@ibp.cz

Keywords: Human genome, DNA sequence, Non-B-DNA, Triplex, Bioconductor, Repetitive sequences, Mobile DNA, Lexicographically minimal rotation

Abstract: Eukaryotic genomes are rich in sequences capable of forming non-B DNA structures. These structures are expected to play important roles in natural regulatory processes at levels above those of individual genes, such as whole genome dynamics or chromatin organization, as well as in processes leading to the loss of these functions, such as cancer development. Recently, a number of authors have mapped the occurrence of potential quadruplex sequences in the human genome and found them to be associated with promoters. In this paper, we set out to map the distribution and characteristics of potential triplex-forming sequences (PTS) in the human genome sequence. Using the R/Bioconductor package *triplex*, we found these sequences to be excluded from exons, while present mostly in a small number of repetitive sequence classes, especially short sequence tandem repeats (microsatellites), Alu and combined elements, such as SVA. We also introduce a novel way of classifying potential triplex sequences, using a lexicographically minimal rotation of the most frequent k-mer to assign class membership automatically. Members of such classes typically have different propensities to form parallel and antiparallel intramolecular triplexes (H-DNA). We observed an interesting pattern, where the predicted third strands of antiparallel H-DNA were much less likely to contain a deletion than their duplex structural counterpart than were their parallel versions.

1 INTRODUCTION

Eukaryotic genomes are rich in sequences capable of forming non-B DNA structures. Cruciform, slipped, triplex or quadruplex DNA has been recognized as a factor in several important biological processes or functions (Buske et al., 2011) (Bacolla and Wells, 2004). Non-B DNA is often found close to recombination hotspots and is thought to aid genomic instability and evolution (Zhao et al., 2010). The structures it forms have the ability to modulate replication (Dixon et al., 2008), transcription (Rich and Zhang, 2008) or translation (Arora et al., 2008) of DNA/RNA by mechanisms that may have their origins in times when nucleic acids dominated all life processes. These structures are expected to play important roles in natural regulatory processes at levels above those of individual genes, such as whole genome dynamics or chromatin organization (Sarkies et al., 2012) (Maizels and Gray, 2013), as well as in processes leading to the loss of these functions, such as cancer development.

For example, a recent study has shown that an interplay between G4-quadruplexes, FANCJ protein and DNA replication in cells influences the formation of euchromatin versus heterochromatin after the replication stage (Schwab et al., 2013). Ability to form H-DNA is often associated with recombinational and mutational hotspots in the human genome (Akman et al., 1991).

Recently, a number of authors have mapped the occurrence of potential quadruplex sequences in the human genome and other eukaryotic genomes. They found them to be associated with promoters and certain classes of repeat elements (Savage et al., 2013) (Lexa et al., 2013). Possibilities of searching for non-B DNA (Cer et al., 2011) and specifically triplex/H-DNA exist as well (Buske et al., 2012) (Hon et al., 2013). In this paper, we map the distribution and characteristics of potential triplex-forming sequences (PTS) in human genomic DNA as detected/predicted using the R/Bioconductor package *triplex*.

2 SOFTWARE AND METHODS

To analyze the human genome, or other sequence sets, we employed the R/Bioconductor framework, which has now matured to the point, where we can use R to represent biological sequences, search these sequences, represent the search results, analyze them statistically and visualize the results of the searches and the statistical analysis. All this can be done with relatively straightforward scripts, using a handful of well integrated R/Bioconductor software packages (Lawrence et al., 2013).

2.1 R/Bioconductor packages used in this study

Biostrings String objects representing biological sequences, and matching algorithms (Pages et al., 2013)

BSgenome Infrastructure for Biostrings-based genome data packages (Pages, 2013)

BSgenome.Hsapiens.UCSC.hg19 Homo sapiens (Human) full genome (UCSC version hg19)

biomaRt Interface to BioMart databases (e.g. Ensembl, COSMIC ,Wormbase and Gramene) (Durinck et al., 2009)

triplex Search and visualize intramolecular triplex-forming sequences in DNA (Hon et al., 2013)

GenomicRanges Representation and manipulation of genomic intervals (Aboyoun et al., 2013)

2.2 General triplex detection pipeline

All types of potential triplexes (parallel and antiparallel) were identified in the human genome using the Bioconductor triplex package (Hon et al., 2013). We used the unmasked sequence from BSGenome.Hsapiens.UCSC.hg19 package in all analyses. Only potential triplexes with P value less than or equal 0.05 were considered for further analysis. A GFF file with all the identified potential triplexes is available at `http://fi.muni.cz/~lexa/triplex/hsapiens_pts.gff`

2.2.1 Analysis of coding and non-coding regions

Information about genes was obtained from the Ensembl database using its Biomart interface. Only coding genes at chromosomes 1-22, X and Y were considered (roughly 20k genes) and only their coding transcripts were selected for analysis (roughly 80k transcripts). Data about exons of selected transcripts were downloaded from Ensembl database and used for identification of promoters, introns, coding regions (CDS), 5'UTR, 3'UTR and intergenic regions. All this information was stored as individual tracks (GRanges objects). For the purpose of this study, promoters were defined as 1000 bp regions upstream of the coding sequence (flanking the 5' end). Intergenic regions were identified as a complement to coding transcripts supplemented with promoters. In the next step we found the overlaps between triplexes and all prepared tracks. If a given triplex fell into more than one type of region (e.g. triplex is part of CDS and intron simultaneously) the triplex was counted in each overlapping region. Finally, we compared the results with numbers expected if positioning of potential triplexes was random. The expected values can be calculated from the percentage of genome covered by a certain type of region (see equations 1-5 below).

2.2.2 Analysis of regions composed of repeats

Information about different types of repeating sequences was obtained from the UCSC Table Browser, specifically from the Repeat Masker track (Karolchik et al., 2004). Data records were organized into 26 classes and 56 families covering both genes (coding and non-coding) and intergenic regions. At first, we analysed the number of potential triplexes in regions with and without repeats in genes and intergenic areas. As the majority of PTS were located in regions with repeats, we performed a detailed study of PTS overlapping individual repeat classes and families. In this experiment we focused on the number of repeats (in a given class or family) containing at least one triplex. The measured values were compared with numbers expected to be obtained at random.

We were also interested in potential triplexes occurring in close proximity to repeats. Therefore we extended all repeat regions with flanking areas (100bps at both ends) and repeated the analysis including these expanded areas.

2.2.3 Calculation of expected values

Expected values were calculated as the number of repeats (of given class/family) that would contain at least one triplex by random choice. This calculation consists of the following steps:

1. Number of triplexes N_{TrRep} that would fall into a given class or family by random is calculated using equation 1.

$$N_{TrRep} = \frac{\sum_{rep \in class} len(rep)}{len(genome)} \tag{1}$$

2. For each repeat rep_{sel} of a given class/family:

   (a) The probability that a randomily selected triplex is placed ouside of a given repeat is calculated using equation 2.

$$P_{RepComp} = \frac{\left(\sum_{rep \in class} len(rep)\right) - len(rep_{sel})}{\sum_{rep \in class} len(rep)} \tag{2}$$

   (b) Next, the probability that all triplexes of a given class/familly are placed outside of a given repeat is calculated using equation 3.

$$P_{RepCompAll} = (P_{RepComp})^{N_{TrRep}} \tag{3}$$

   (c) Finaly, the probility that at least one triplex falls into a given repeat is calculated using equation 4.

$$P_{Rep} = 1 - P_{RepCompAll} \tag{4}$$

3. The overall number of repeats that would contain at least one triplex by random choice is calculated as a sum of all probabilities calculated in the previous step (see equation 5).

$$N_{Rep} = \sum_{Rep \in class} P_{Rep} \tag{5}$$

Please note that in equations 1 and 2 we use a sum for expression of area occupied by repeats. In fact the real calculation is slightly more complex because the overlapping repeat regions have to be considered as well. Concretely, if two repeats of the same class or family overlap each other then the overlapping part is counted only once in that sum.

Expected values for repeats supplemented with flanking areas are calculated analogically.

2.2.4 Analysis of non-coding genes

Data about non-coding genes were obtained from the Ensembl database using the Biomart interface (roughly 40k genes). All genes were split into 26 categories based on their biotype (e.g. lincRNA, pseudogene, miRNA, scRNA, etc.) The same type of analysis, which was performed for repetitive sequences, was applied for non-coding genes as well.

All types of experiments were applied separately on a set of parallel, antiparallel and all types of triplexes (parallel and antiparallel together).

2.3 Classification of potential triplex sequences

H-DNA often forms in sequences that contain simple repetitions. The type of triplex that forms (e.g. parallel or antiparallel) often depends on the particular kind of repeat present. We therefore decided to classify the identified triplexes by the prevailing k-mers present in their sequences. Although different values of k will serve the purpose of classification, we selected $k = 6$ as a value which is not too high since high values of k would produce a large number of categories that would be difficult to follow. Because $k = 6$ is also the lowest number that can capture well both, periods of 2 and 3, we chose this value for all calculations in this study.

The prevailing k-mers for identical classes of sequences will sometimes differ, because in a periodic sequence, all rotations of the repeated sequence monomer will have similar chance of becoming the most prevalent k-mer. The precise result will depend on subtle changes in the sequence. For example, for $k = 2$, the sequence *GCGCGCCGCGC* will have 4 CG dinucleotides and 5 GC dinucleotides, we would therefore label this sequence as ”GC”. A very similar sequence *CGCGCGGCGCG* with one $C- > G$ substitution and one nucleotide moved from the end to the front has 5 CGs and 4GCs and would therefore be labeled as ”CG”.

In situations were several rotations of a string may represent the same feature, we can deterministicaly choose one of the variants (rotations) to represent all of them. One way of choosing the representative string (hexamer) is to choose the one that comes first in lexicographical order. We therefore propose to classify and label the sequences with the lexicographically minimal rotation of the most prevalent k-mer. This way the DNA sequence gets labeled by the most prevalent sequence motif, regardless of its exact distribution. In R, the classification into labelled classes was achieved by the following R code:

```
triplex_class <- function(x,k){
s <- as.character(x)
n <- nchar(s)
res <-
 names(sort(table(substring(s,1:(n-k+1),k:n))
 ,decreasing=TRUE)[1])
# the lexicographically minimal rotation
res2 <- paste(res,res,sep="")
sort(substring(res2,1:6,6:11))[1]
}
```

One can then easily annotate sequences identified by *triplex.search()*, and stored in the variable *tc*, simply by calling

```
sort(table(sapply(tc,triplex_class,6))
,decreasing=TRUE)
```

This novel method of annotation will probably need to be further refined, since in its proposed form it only works well if the value of k used in the analysis is equal to the intrinsic periodicity of the analysed sequence. Using hexamers succesfully captures periods of 1,2,3 and 6 but may give fragmented results for other periods. We presently solve this by manually choosing a class name based on an inspection of the hexamer, such as CTT/GAA (see Table 1).

2.4 Insertions or deletions in potential triplex sequences

The output of *triplex.alignment()* contains the designation of individual H-DNA strands. They are labelled as *plus*, *minus*, *par*+, *par*−, *apar*+, *apar*− and *loop*. We used the script *count_chars.R* to determine the frequency of insertions/deletions (symbol ”-”) or other symbols in the aligned DNA strands. We expected different tolerance for insertions/deletions between strands, because the properties of the original duplex (strands *plus* and *minus* may be quite different from the properties of the third DNA strand attached via Hoogsteen or reverse Hoogsteen bonds. The counted insertions were compared to the total length of each type of strand (regular expression ”.” for any symbol).

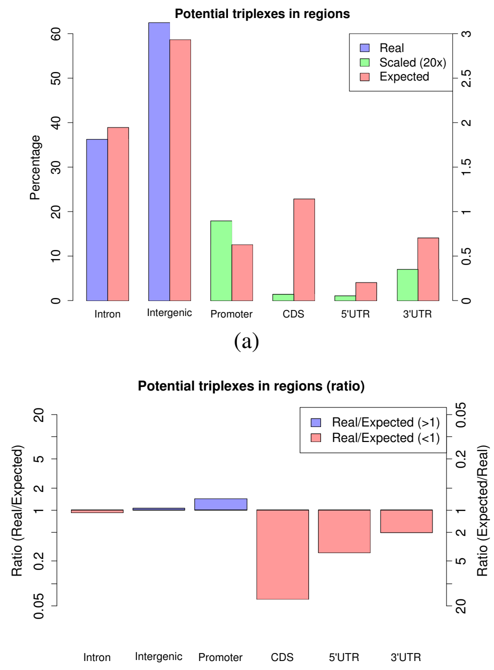
3 RESULTS

A series of calculations were carried out as described in section 2 to understand the distribution of PTS in the human genome and determine the properties of these sequences.

3.1 Distribution of potential triplex sequences in the human genome

To obtain an overall picture of how potential triplex sequences (PTS) are distributed in the human genome, we first compared the PTS positions with the positions of general annotated genome features, such as protein-coding sequences, promoters, introns and intergenic regions. There is a clear preference of PTS to be present in promoters or intergenic regions with close-to-predicted content in introns and strong avoidance of exons, including 5’- and 3’-UTRs (Figure 1).

Because the intergenic regions in the human genome are also known to harbor a high number of repetitive sequences (e.g. 10% Alu (SINE), 15% L1 (LINE), 8% LTR retrotransposons) we calculated the number of PTS associated with individual repetitive sequence classes (Figure 2(a)) and families (Figure 2(b)). This analysis shows that only a limited number of repeat classes and families associate with H-DNA more frequently than expected from genome averages.



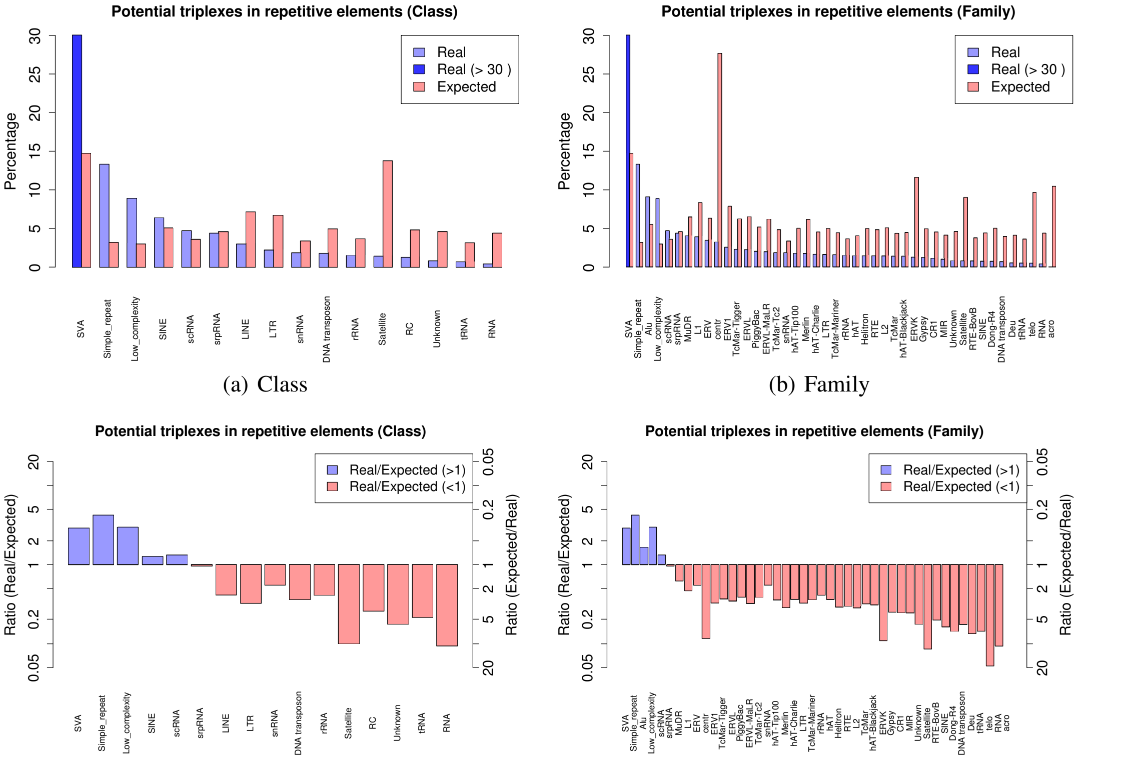
(a)

(b)

Figure 1: Distribution of potential triplex sequences (PTS) in genes and intergenic regions. The figure (a) shows the percentage of all elements in a given region. The four bars on the right side, namely Promotor, CDS, 5’UTR and 3’UTR, were zoomed in 20× and supplemented with its own axis. The figure (b) shows the ratio between real and expected counts. The expected number of triplexes was estimated from triplex density and the overall length of sequence in a given category.

To allow better comparison across families and classes that would not depend on repeat size and frequency of occurrence, we calculated the ratio between real and expected counts. In both types of analysis, there is a strong enrichment of PTS sequences in low complexity sequences and simple repeats (Figure 2(c)). Such findings are compatible with the limited number of nucleotide triplets found in stable H-DNA (Soyfer and Potaman, 1995) (Lexa et al., 2011) and the general requirement for polypurine and polypyrimidine tracts in triplexes. We also found SVA and to a limited extent also SINE and scRNA elements to have above average PTS association (Figure 2(c)). Of these, specifically SVA and Alu sequences showed above average association (Figure 2(d)).

We also looked whether there was a difference be-



(a) Class

(b) Family

(c) Class ratio

(d) Family ratio

Figure 2: Association of potential triplex sequences (PTS) with different families of repetitive sequences identified by RepeatMasker. The figures (a) and (b) show the percentage of all elements in a given family harboring at least one PTS. The figures (c) and (d) show ratio between real and expected counts test The expected number of triplexes was estimated from triplex density and the overall length of sequences of a given family.

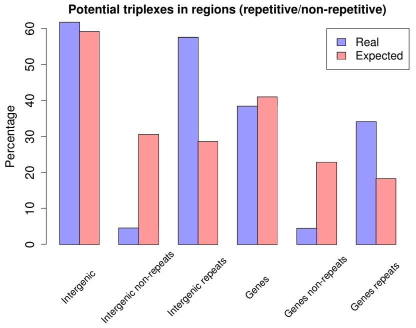


Figure 3: Occurrence of repeat and non-repeat-associated potential triplex sequences (PTS) in different regions of the human genome. Data shown as percentage of all PTS in the genome. The expected percentages of triplexes were estimated from average triplex density and the overall length of sequences of a given genomic region.

tween PTS sequences found in the different regions of the human genome. While the frequency of PTS in intergenic regions was slightly higher than in genes, were they prevailed in introns, they were equally associated with repeats in both parts, introns and intergenic (Figure 3).

3.2 Classification of potential triplexes by sequence composition

To classify the detected PTS, we counted all nucleotide hexamers in their sequence and determined the lexicographically minimal rotation of the prevalent hexamer as described in section 2. This analysis revealed the presence of six main categories of PTS in the human genome (Table 1). In the order of prevalence, these were classes labelled by us as T/A (45.8%), CT/GA (20.6%), CTT/GAA (14.6%), CCT/GGA (13.1%), C/G (3.6%), CA/GT (1.6%) and TA/TA (0.5%). The remaining PTS constituted only 0.3% of detected sequences.

Table 1: Classification of PTS by sequence composition and the occurrence of different composition classes in the human genome.

Hexamer	Count	[%]	Class/Composition
TTTTTT	4360	16.6	T/A
AAAAAA	4346	16.5	T/A
AAAAAG	1993	7.6	T/A
CTCTTT	1564	5.9	CT/GA
CTCTCT	1470	5.6	CT/GA
AGAGAG	1444	5.5	CT/GA
CTTTTT	1337	5.1	T/A
CCCCTT	992	3.8	CCT/GGA
AAAAGG	954	3.6	CTT/GAA
AAAGAG	950	3.6	CT/GA
CCCCCT	624	2.4	C/G
AGAGGG	568	2.2	CCT/GGA
AAGGGG	528	2.0	CCT/GGA
CCTCCT	520	2.0	CCT/GGA
CCCTCT	508	1.9	CCT/GGA
CCTTTT	480	1.8	CTT/GAA
CTTCTT	439	1.7	CTT/GAA
CCTTCT	417	1.6	CTT/GAA
AAGAAG	410	1.6	CTT/GAA
AAGAGG	345	1.3	CTT/GAA
AGGAGG	326	1.2	CCT/GGA
AAGGAG	314	1.2	CTT/GAA
AGGGGG	311	1.2	C/G
CCTCTT	285	1.1	CTT/GAA
GTGTGT	225	0.9	CA/GT
ACACAC	187	0.7	CA/GT
ATATAT	144	0.5	TA/TA
AAAGGG	109	0.4	CTT/GAA
CCCTTT	85	0.3	CTT/GAA
OTHER	68	0.3	-

3.3 Tolerance of different potential triplex classes and aligned DNA strands for insertions

We hypothesized that triplex sequences have asymmetric tolerance for insertions/deletions. Because the third strand binds to a DNA duplex in its major grove with lower stringency than seen in Watson-Crick basepaired duplex, the third strand may be able to accept insertions, but not deletions when aligned to the duplex with *triplex.alignment()*. Although the ability of H-DNA to accept mismatches, let alone indels, is highly questionable, we still carried out this calculation, counting the occurrence of the ”-” symbol in different strands and counting the overall length (Table 2). We found that only 0.75% of positions in PTS were insertions/deletions in PTS scoring 25 or more. Moreover, when we compared the occurence of deletions in the different types of PTS third strand to the frequencies observed in the duplex at various score thresholds, we found the percentage to be lower in parallel strands (0.87-0.97%) and much lower in antiparallel strands (0.16-0.22%) (Table 3). These data appear to support our hypothesis of asymmetrical insertion/deletion distribution among DNA strands in potential triplexes.

4 DISCUSSION

We have taken a closer look at the output of the R/Bioconductor triplex search package when ran against the human genome DNA sequence. In terms of search results, we were interested to see the different categories of human sequences that associate with potential intramolecular triplexes. The slight over-representation of PTS in non-coding sequences and clear absence from coding sequences seen in Figure 1 led us to focus on intergenic DNA, promoters and introns in more detail (Figure 2(a), 2(b)). H-DNA has been found in promoters of genes involved in disease (Bissler, 2007) and cell signalling and communication (Bacolla et al., 2006).

There is a common theme to the majority of PTS occurrences we observed in human DNA. Inspection of Figure 2(b)-2(d) reveals the presence of PTS in or near Alu, scRNA and simple repeat or low complexity sequences. Alu sequences are short non-autonomous retrotransposons (SINE) driven by the L1 LINE element protein machinery (Dewannieux et al., 2003) thought to have emerged in primate as duplication descendants of 7SL sc RNA (Kriegs et al., 2007). SVA repeats, which contained more then twice the number of PTS than expected by chance are also strongly associated with PTS. Perhaps not surprisingly, even SVA elements are evolutionarily related to SINE and Alu sequences. Their sequence is chimeric and contains two sequences of SINE origin separated by a variable number tandem repeat (Savage et al., 2013). According to our study, a large proportion of PTS in the human genome can therefore be directly attributed to the proliferation of SINE elements, especially Alu.

Upon first inspection, it becomes clear that most of the above-mentioned associations are caused by the presence of the polyA tail in SINE elements. Because the poly-A tail is mainly described as a feature circumventing the problematic polyadenylation in RNA polymerase III transcripts (Roy-Engel, 2012), there is a possibility that these sequences do not form any functionally or evolutionarily meaningful DNA structures, such as H-DNA. On closer inspection, however, we notice that the same classes of repeats are also enriched for other PTS sequences, raising the possibility that triplex formation plays a biological role in the repeat life cycles also at the DNA level. This could also mean a dual role for the Alu poly-A tail. For example, (Dewannieux and Heidmann, 2005) mention a 15-50 nucleotide range for increasing effect of the poly-A tail, a range that also coincides with cited oligonucleotide lengths for successful H-DNA formation (Buske et al., 2011). (CT)$_n$ tandem repeats have also been implicated in tandem array maintenance (Bailey et al., 2013), the mechanism and its dependence on triplex formation is, however, presently unknown. (Brereton et al., 1993) showed that the A-rich sequence in a human Alu element can form an intramolecular triplex *in vitro*.

Given the presence of PTS in Alu and SVA repeats in human, that have evolved as dimers (the former) and dimer of dimers (the latter) of ancient RNA, there is a possibility for intramolecular triplexes to aid the recombination processes leading to chimeric sequences. There are indeed many reports of H-DNA occurrence near recombination hotspots (Napierala et al., 2004).

Because of the high Alu content of the human genome, the presence of PTS in Alu elements biased our estimates of expected PTS association frequency with other elements. Upon subtracting these from our results, several other classes get into the "above-expected" occurrence territory, namely the L1 retrotransposon and MuDR DNA transposon as well as snRNA which often contains a (CT)n dinucleotide tandem repeat.

We have also noticed a high occurrence of PTS in the miRNA class of RNAs (data not shown). Kanak and colleagues (Kanak et al., 2010) recently reported the discovery of a set of miRNA sequences that could form triplexes at HIV target sites and suppress its retroviral activity. An increased presence of PTS sequences has recently been reported in the 5' and 3'-UTR of plant retroelements, probably analogous to the reported 3'-UTR HIV regulatory region.

Probably the second most typical location for PTS in our study were the promoters of genes. The formation of special DNA structures at sequences such as PTS studied in this paper may create structurally distinct features providing possibilities for specific DNA-binding proteins to recognize locations in the genome for gene regulation or chromatin organization. For example, triplex DNA has been found to be incompatible with nucleosome formation and may act as a nucleosome barrier (Westin et al., 1995). They are often found near recombination and mutation hotspots (Napierala et al., 2004)(Akman et al., 1991). This may be related to the inevitability of single-stranded DNA stretches at or near the triplexes. Association of PTS with certain types of repeat elements could not only suggest a possible function in the repeat "life cycle" but also a possible positive selection for repeats with such association, if the presence of triplexes was required at several locations of the host genome.

Among the typical hexamers found in triplexes, we identified a minor group wih prevailing CA/GT dinucleotide repeats. Although this combination does not meet the often cited requirement for homopurine and homopyrimidine tracts in H-DNA, it may actually form intramolecular triplexes in combination with other base triplets, as observed by (Gowers and Fox, 1998). Our extremely low counts (Table 1) seem to support the notion that if G.T:A and T.A:T triplets occur in triplexes, they are most likely to be mixed with other nucleotide combinations.

5 CONCLUSIONS

In this paper we examined the types of sequences that can be identified in the human genome DNA sequence with triplex DNA detection software, namely the R/Bioconductor package triplex-1.0.10 and its *triplex.search()* function. The presented results examine the usability of the software for genome studies as well as some basic properties of the identified potential triplex sequences (PTS). We found that most of the triplex-forming potential of the human genome is concentrated in simple repeats and flanking regions of repetitive and other genome elements descending from 7SL RNA, especially Alu and SVA repeats. We also found potential triplex-forming sequences in the miRNA class of RNA genes. Alu elements are known to contain or flank adenine homonucleotide tracts which replace polyadenylation of its RNA, but could also carry out a DNA-based function involving H-DNA formation.

We propose a computational rule to automatically classify triplex-forming sequences according to the most prevalent *k*-mer present in their sequence. For unambiguity, we include the search for a lexicographically minimal rotation before assigning the name. After applying this principle we see that the majority of human PTS fall into four main classes based on their nucleotide composition (T/A - 45.8%; CT/GA - 20.6%; CTT/GAA - 14.6% and CCT/GGA - 13.1%). We also characterized the detected PTS based on deletions found in alignments of the third triplex strand to the DNA duplex, sowing that deletions are present less frequently in the third strand, especially in antiparallel PTS.

In terms of biological relevance, our studies of PTS suggest they are positioned non-randomly in the genome, their sequences fall into a small number of distinct classes and some of them are associated with specific types of repeats. Their strand bias for insertions or deletions suggests that these sequences may indeed form the predicted structures. In future it would be desireable to single out specific combination of repeat types and PTS classes, prove the existence of triplex formation in each case and systematically search for proteins that could interact with such structures and provide a more precise clue to their specific biological function.

ACKNOWLEDGEMENTS

This work was supported by research grants from the European Social Fund (CZ.1.07/2.3.00/20.0189) to M.L., 13-36108S (Grant Agency of Science of CR) to M.B., by IT4Innovations project, reg. no. CZ.1.05/1.1.00/02.0070 funded by the EU Operational Programme, by MSMT Grant No.0021630528 and by BUT grant FIT-S-11-1.

REFERENCES

Aboyoun, P., Pages, H., and Lawrence, M. (2013). Genomicranges: Representation and manipulation of genomic intervals. Technical Report R package version 1.10.7.

Akman, S. A., Lingeman, R. G., Doroshow, J. H., and Smith, S. S. (1991). Quadruplex dna formation in a region of the trna gene supf associated with hydrogen peroxide mediated mutations. *Biochemistry*, 30(35):8648–8653.

Arora, A., Dutkiewicz, M., and Scaria, V. (2008). Inhibition of translation in living eukaryotic cells by an rna g-quadruplex motif. *RNA*, 14:1290–1296.

Bacolla, A. and Wells, R. (2004). Non-b dna conformations, genomic rearrangements, and human disease. *Journal of Biological Chemistry*, 279:47411–47414.

Bacolla, A., Wojciechowska, M., Kosmider, B., Larson, J. E., and Wells, R. D. (2006). The involvement of non-b dna structures in gross chromosomal rearrangements. *DNA Repair*, 5:1161–1170.

Bailey, A. D., Pavelitz, T., and Weiner, A. M. (2013). The microsatellite sequence (ct)n.(ga)n promotes stable chromosomal integration of large tandem arrays of functional human u2 small nuclear rna genes. *Molecular and Cellular Biology*, 18(4):2262–2271.

Bissler, J. J. (2007). Triplex dna and human disease. *Frontiers in Bioscience*, 12:4536–4546.

Brereton, H., Firgaira, F., and Turner, D. (1993). Origins of polymorphism at a polypurine hypervariable locus. *Nucleic Acids Research*, 21(11):2563–2569.

Buske, F. A., Bauer, D. C., Mattick, J. S., and Bailey, T. L. (2012). Triplexator: detecting nucleic acid triple helices in genomic and transcriptomic. *Genome Research*, 22(7):1372–1381.

Buske, F. A., Mattick, J. S., and Bailey, T. L. (2011). Potential in vivo roles of nucleic acid triple-helices. *RNA Biology*, 8(3):427–439.

Cer, R. Z., Bruce, K. H., Mudunuri, U. S., Yi, M., Volfovsky, N., Luke, B. T., Bacolla, A., Collins, J. R., and Stephens, R. M. (2011). Non-b db: a database of predicted non-b dna-forming motifs in mammalian genomes. *Nucleic Acids Research*, 39(Database issue):D383–D391.

Dewannieux, M., Esnault, C., and Heidmann, T. (2003). Line-mediated retrotransposition of marked alu sequences. *Nature Genetics*, 35:41–48.

Dewannieux, M. and Heidmann, T. (2005). Role of poly(a) tail length in alu retrotransposition. *Genomics*, 86(3):378–381.

Dixon, B., Lu, L., Chu, A., and Bissler, J. (2008). Recq and recg helicases have distinct roles in maintaining the stability of polypurine.polypyrimidine sequences. *Mutation Research*, 643:20–28.

Durinck, S., Spellman, P., Birney, E., and Huber, W. (2009). Mapping identifiers for the integration of genomic datasets with the r/bioconductor package biomart. *Nature Protocols*, 4:1184–1191.

Gowers, D. and Fox, K. (1998). Triple helix formation at (at)n adjacent to an oligopurine tract. *Nucleic Acids Research*, 26(16):3626–3633.

Hon, J., Martinek, T., Rajdl, K., and Lexa, M. (2013). Triplex: an r/bioconductor package for identification and visualization of potential intramolecular triplex patterns in dna sequences. *Bioinformatics*, 29(15):1900–1901.

Kanak, M., Alseiari, M., Balasubramanian, P., Addanki, K., Aggarwal, M., Noorali, S., Kalsum, A., Mahalingam, K., Pace, G., Panasik, N., and Bagasra, O. (2010). Triplex-forming micrornas form stable complexes with hiv-1 provirus and inhibit its replication. *Applied Immunohistochemistry and Molecular Morphology*, 18(6):532–545.

Karolchik, D., Hinrichs, A. S., Furey, T. S., Roskin, K. M., Sugnet, C. W., Haussler, D., and Kent, W. J. (2004). Ucsc table browser data retrieval tool. *Nucleic Acids Research*, 32(Database issue):D493–D496.

Kriegs, J. O., Churakov, G., Jurka, J., Brosius, J., and Schmitz, J. (2007). Evolutionary history of 7sl rna-derived sines in supraprimates. *Trends in Genetics*, 23(4):158–161.

Lawrence, M., Huber, W., Pags, H., Aboyoun, P., Carlson, M., Gentleman, R., Morgan, M. T., and Carey, V. J. (2013). Software for computing and annotating genomic ranges. *PLoS Computational Biology*, 9(8):e1003118.

Lexa, M., Kejnovsky, E., Steflova, P., Konvalinova, H., Vorlickova, M., and Vyskot, B. (2013). Quadruplex-forming sequences occupy discrete regions inside plant ltr retrotransposons. *Nucleic Acids Research*, page 10.1093/nar/gkt893 (ePub).

Lexa, M., Martinek, T., Burgetova, I., Kopecek, D., and Brazdova, M. (2011). A dynamic programming algorithm for identification of triplex-forming sequences. *Bioinformatics*, 27(18):2510–2517.

Maizels, N. and Gray, L. (2013). The g4 genome. *PLoS Genetics*, 9(4):e1003468.

Napierala, M., Dere, R., Vetcher, A. A., and Wells, R. D. (2004). Dna replication repair and recombination: Structure-dependent recombination hotspot activity of gaattc sequences from intron 1 of the friedreich's ataxia gene. *The Journal of Biological Chemistry*, 279:6444–6454.

Pages, H. (2013). Bsgenome: Infrastructure for biostrings-based genome data packages. Technical Report R package version 1.26.1.

Pages, H., Aboyoun, P., Gentleman, R., and DebRoy, S. (2013). Biostrings: String objects representing biological sequences, and matching algorithms. Technical Report R package version 2.26.3.

Rich, A. and Zhang, S. (2008). Timeline: Z-dna: the long road to biological function. *Nature Reviews Genetics*, 4:566–572.

Roy-Engel, A. M. (2012). A tale of an a-tail. the lifeline of a sine. *Mobile Genetic Elements*, 2(6):282–286.

Sarkies, P., Murat, P., Phillips, L., Patel, K., Balasubramanian, S., and Sale, J. (2012). Fancj coordinates two pathways that maintain epigenetic stability at g-quadruplex dna. *Nucleic Acids Research*, 40(4):1485–1498.

Savage, A. L., Bubb, V. J., Breen, G., and Quinn, J. P. (2013). Characterisation of the potential function of sva retrotransposons to modulate gene expression patterns. *BMC Evolutionary Biology*, 13(101).

Schwab, R. A., Nieminuszczy, J., Shin-ya, K., and Niedzwiedz, W. (2013). Fancj lets chromatin stay true. *Journal of Cell Biology*, 201:33–48.

Soyfer, V. and Potaman, V. (1995). *Triple-helical nucleic acids.* Springer-Verlag, Heidelberg.

Westin, L., Blomquist, P., and Milligan, J. F. e. a. (1995). Triple helix dna alters nucleosomal histone-dna interactions and acts as a nucleosome barrier. *Nucleic Acids Reserch*, 23:2184–2191.

Zhao, J., Bacolla, A., Wang, G., and Vasquez, K. (2010). Non-b dna structure-induced genetic instability and evolution. *Cellular and Molecular Life Sciences*, 67(1):43–62.

Table 2: The number of deletions counted in different strands of PTS in the human genome DNA sequence and the total number of PTS strands examined.

	Number of deletions					Number of PTS strands of a given type				
Score	duplex	para+	para-	anti+	anti-	duplex	para+	para-	anti+	anti-
> 25	7953	3053	2931	202	170	956452	378942	357804	109834	109872
> 35	7443	2934	2834	116	105	853345	365429	345811	70800	71305
> 50	3972	1630	1624	8	10	425395	206228	200898	8985	9284
> 70	1001	442	437	0	0	148244	72251	73930	897	1166

Table 3: The frequency and relative occurrence of deletions in DNA strands of different classes in human PTS.

	100*deletions/strand					relative to duplex	
Score	duplex [%]	para+ [%]	para- [%]	anti+ [%]	anti- [%]	para/duplex	anti/duplex
> 25	0.83	0.81	0.82	0.18	0.15	0.98	0.19
> 35	0.87	0.80	0.82	0.16	0.15	0.93	0.17
> 50	0.93	0.79	0.81	0.09	0.11	0.86	0.10
> 70	0.68	0.61	0.59	0.00	0.00	0.88	0.00

A.5 Paper V

p53 Specifically Binds Triplex DNA In Vitro and in Cells

PLOS | ONE

RESEARCH ARTICLE

p53 Specifically Binds Triplex DNA *In Vitro* and in Cells

Marie Brázdová[1]*, Vlastimil Tichý[1], Robert Helma[1], Pavla Bažantová[1], Alena Polášková[1], Aneta Krejčí[2], Marek Petr[1], Lucie Navrátilová[1], Olga Tichá[1], Karel Nejedlý[1], Martin L. Bennink[3], Vinod Subramaniam[3], Zuzana Bábková[2], Tomáš Martínek[4], Matej Lexa[5], Matej Adámik[1]

1 Department of Biophysical Chemistry and Molecular Oncology, Institute of Biophysics, Academy of Sciences of the Czech Republic v.v.i., Brno, Czech Republic, **2** Department of Molecular Biology and Pharmaceutical Biotechnology, Faculty of Pharmacy, University of Veterinary and Pharmaceutical Sciences Brno, Brno, Czech Republic, **3** Biophysical Engineering Group, Faculty of Science and Technology, University of Twente, Enschede, The Netherlands, **4** Department of Computer Systems, Faculty of Information Technology, Brno University of Technology, Brno, Czech Republic, **5** Department of Information Technologies, Faculty of Informatics, Masaryk University, Brno, Czech Republic

* maruska@ibp.cz

Abstract

Triplex DNA is implicated in a wide range of biological activities, including regulation of gene expression and genomic instability leading to cancer. The tumor suppressor p53 is a central regulator of cell fate in response to different type of insults. Sequence and structure specific modes of DNA recognition are core attributes of the p53 protein. The focus of this work is the structure-specific binding of p53 to DNA containing triplex-forming sequences *in vitro* and in cells and the effect on p53-driven transcription. This is the first DNA binding study of full-length p53 and its deletion variants to both intermolecular and intramolecular T.A.*T* triplexes. We demonstrate that the interaction of p53 with intermolecular T.A.*T* triplex is comparable to the recognition of CTG-hairpin non-B DNA structure. Using deletion mutants we determined the C-terminal DNA binding domain of p53 to be crucial for triplex recognition. Furthermore, strong p53 recognition of intramolecular T.A.*T* triplexes (H-DNA), stabilized by negative superhelicity in plasmid DNA, was detected by competition and immunoprecipitation experiments, and visualized by AFM. Moreover, chromatin immunoprecipitation revealed p53 binding T.A.*T* forming sequence *in vivo*. Enhanced reporter transactivation by p53 on insertion of triplex forming sequence into plasmid with p53 consensus sequence was observed by luciferase reporter assays. *In-silico* scan of human regulatory regions for the simultaneous presence of both consensus sequence and T.A.*T* motifs identified a set of candidate p53 target genes and p53-dependent activation of several of them (*ABCG5*, *ENOX1*, *INSR*, *MCC*, *NFAT5*) was confirmed by RT-qPCR. Our results show that T.A.*T* triplex comprises a new class of p53 binding sites targeted by p53 in a DNA structure-dependent mode *in vitro* and in cells. The contribution of p53 DNA structure-dependent binding to the regulation of transcription is discussed.

OPEN ACCESS

Citation: Brázdová M, Tichý V, Helma R, Bažantová P, Polášková A, Krejčí A, et al. (2016) p53 Specifically Binds Triplex DNA *In Vitro* and in Cells. PLoS ONE 11(12): e0167439. doi:10.1371/journal.pone.0167439

Editor: Sergey Korolev, Saint Louis University, UNITED STATES

Received: July 23, 2016

Accepted: November 14, 2016

Published: December 1, 2016

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: This work was supported by the Czech Science Foundation (13-36108S and 204/06/P369 to MB, 15-02891S to ML), University of Veterinary and Pharmaceutical Sciences Brno (IGA VFU Brno 316/2016/FaF to AK), IT4Innovations excellence in science project (IT4I XS – LQ1602 to TM) and by the Czech Academy of Sciences (RVO68081707 to VT, MB, RH, PB, AP, MP, LN, MA).

Competing Interests: The authors have declared that no competing interests exist.

Introduction

Tumor suppressor p53 contains two DNA binding domains. The central (core) domain (amino acids ~100 to ~300) is evolutionarily highly conserved and is essential for p53 sequence-specific binding to promoters of p53 target genes that take part in cell cycle regulation, apoptosis and DNA repair [1]. The p53 consensus sequence (CON) has been originally defined as two copies of the sequence 5´-PuPuPuC(A/T)(T/A)GPyPyPy-3´ separated by 0–13 bp [2]. The core domain also binds in non-sequence-specific manner to single- and double-stranded DNA, preferentially interacting with internal regions of single-stranded (ss) DNA [3], three-stranded DNA substrates mimicking early recombination intermediates [4], insertion/deletion mismatches [5] and DNA cruciform stabilized by DNA superhelicity [6]. The C-terminal part of the protein contains a flexible linker (amino acids ~300 to ~325), a tetramerization domain (amino acids ~325–356) and a basic C-terminal DNA binding domain (CTDBD, aa 363–382). The ability of the C-terminus to bind single-stranded gaps in double-stranded (ds) DNA [7], cisplatin-modified DNA [8], hemicatenated DNA loops [9] and superhelical DNA (scDNA [10, 11]) has been described. There is a growing amount of data suggesting that p53 interactions with different DNA targets represent a complex network involving contributions from both DNA binding domains reviewed in [12]. Recently, we have shown that the human telomeric G-quadruplexes are recognized by full length p53 protein and both DNA-binding domains take part in this interaction [13].

The triple-helical (triplex) DNA adopts a structure characterized by a third pyrimidine-rich or purine-rich DNA strand located within the major groove of a homopurine/homopyrimidine stretch of duplex DNA [14–16]. Stable interaction of the third strand is achieved through either specific Hoogsteen or reverse Hoogsteen hydrogen bonding with the homopurine strand of the duplex. Preferred base triplets include T.A.*T* and C.G.*C* in the pyrimidine motif and C.G.*G* and T.A.*A* in the purine motif. Triplexes can be either intermolecular, where the third strand originates from a separate DNA molecule, or intramolecular (named also H-DNA), where the third strand originates from the same DNA molecule as its duplex acceptor [15, 16]. Naturally occurring sequences capable of forming intramolecular triplex are found in human genome as frequently as 1 in every 50000 bp [17] and are enriched in introns and promoters [18, 19]. Intramolecular triplexes are postulated to occur *in vivo* under suitable conditions (such as sufficiently high negative superhelical stress) and their involvement has been implicated in several cellular processes, including transcription, replication and recombination [15, 16]. The triplex target sequence for formation of intermolecular DNA triplexes is even more abundant, on average one unique triplex target sequence every 1366 bases [20]. Intermolecular triplexes are widely recognized as potential tools for different genetic manipulations including gene regulation and mutagenesis [21, 22]. So far, only a few proteins recognizing triplexes of pyrimidine type are known [23–26]. The importance of triplex DNA for the occurrence of some breakpoint hotspots in cancer has also been hypothesized [27]. Despite the correlation between genomic instability and formation of triplex DNA, the function of proteins that recognize these structures is still poorly understood. Several DNA repair proteins have been shown to bind triplex DNA [23].

Negative DNA superhelicity is necessary for the formation of intramolecular triplex DNA (H-DNA) and other non-B DNA structures *in vivo* [28]. Observations from our laboratory [11, 29, 30], as well as of others [12, 31] have revealed a clear relationship between the topology of recognized DNA and p53. Both wild-type p53 and mutant p53 proteins have considerable potential to recognize non-B DNA structures. In particular, formation of stem-loop, hairpin or cruciform structures affects p53-DNA interactions [12, 30–33].

In this study, we have analyzed for the first time the interaction of the full-length p53 and its deletion variants to DNA containing triplex-forming sequences *in vitr*o and in cells. We show that p53 protein possessing intact C-terminus exhibits high affinity to intermolecular and intramolecular T.A.*T* triplex DNA. *In-silico* analysis of human promoters for simultaneous presence of consensus sequence and T.A.*T* motifs identified a set of candidate p53 target genes. Possible contribution of DNA triplex-dependent binding of p53 for regulation of their transcription is discussed.

Material and Methods

Oligonucleotides

The sequences of oligonucleotides used in this study are presented in S1 Table, oligonucleotides were synthesized by VWS (Vienna, Austria). Duplex and triplex probes were prepared as previously described [25]. Briefly, intermolecular T.A.*T* triplex (oligo$(dT)_{50}$.oligo$(dA)_{50}$.oligo $(dT)_{50}$) was formed by standard annealing of $(dT)_{50}$ to labeled $(dA)_{50}$ and titration of duplex with $(dT)_{50}$ to molar excess (3–5×) in presence of Mg^{2+} ions in triplex forming buffer (5 mM Tris-HCl, pH 8, 1 mM $MgCl_2$, 300 mM NaCl) at 37°C for 60 min. $CTG_{hairpin}$ and $TA_{hairpin}$ were prepared as described in [32] with labeled lock oligonucleotide (S1 Table).

Recombinant plasmids

Plasmids encoding human p53 proteins pT7-7wtp53 (full length wild type p53,p53, aa 1–393), pET-p53CD (p53CD, aa 94–312), pGEX-2TKp53CT (GST-p53CT, aa 320–393), pGEX-2TKp53T (GST-p53T, aa 363–393) and pGEX-4Tp53CD (GST-p53CD, aa 94–312) were described in [10, 29]. Plasmids with T.A.*T* triplex forming sequences (pBA50 and pPA50) were prepared by cloning of $(dT)_{50}$.$(dA)_{50}$ into the *Eco*RV site of pBluescript SK II- (pBSK, Stratagene) and pPGM1 [34] (S1 Table). Similarly, plasmids for cruciform formation (pBAT34, pPAT34) were prepared by cloning $(dAdT)_{34}$ sequences to the same plasmids, for details see S1 Table. Plasmid pA69 with $(dT)_{69}$.$(dA)_{69}$ (on pUC19 basis [35]) and pUC19 control plasmid were used. Nonspecific competitor (pBSK/*Sma*I) was prepared by *Sma*I restriction enzyme (Takara, Japan) cleavage of pBSK. Plasmids for luciferase reporter assay (pGL3-BSK, pGL3-P1, pGL3-BA50, pGL3-PA50, pGL3-PA20, S1 Table) were constructed by cloning fragments from pBSK derivatives into the *Sma*I/*Xho*I site of the pGL3-promoter (Invitrogen). All plasmids were isolated from bacterial strain TOP10 (Stratagene) and verified by sequencing.

p53 recombinant proteins purification

Full length p53 and isolated DNA binding domains p53CD, p53CT, and p53T (with or without GST tag) were purified according to a protocol described previously [10, 29]. The purity and appropriate size of each protein were analyzed by Coomassie blue staining of 12.5% SDS-PAGE gels (S1A Fig), using bovine serum albumin as a standard.

EMSA in polyacrylamide gels

^{32}P-radiolabeled oligonucleotide probes (1 pmol) were mixed with p53 proteins and incubated in binding buffer (5 mM Tris-HCl, pH 8, 1 mM $MgCl_2$, 0.01% Triton X-100 and 50 mM KCl) in the presence of 5–50 ng pBSK/*Sma*I competitor DNA for 30 min on ice or at 25°C to reach equilibrium. Samples were loaded onto a 4–5% polyacrylamide gel containing 0.5× TB buffer with 2 mM $MgCl_2$. After 1–3 h electrophoresis (at 4–6 V/cm^2) the gels were dried and DNA was detected by autoradiography using Typhoon FLA 9000 (GE Healthcare). Polyclonal rabbit

CM1 and mouse monoclonal (DO1 (aa 20–25), Bp53-10.1 (aa 375–379), PAb421 (aa 371–380) and ICA9 (aa 388–393)) antibodies, kindly provided by Dr. B. Vojtesek, were used in supershift and IP experiments.

ELISA

96-well Immuno Plates (SPL LIFE SCIENCES) were streptavidin (PROSPEC) coated and blocked for unspecific binding by BSA (Sigma). Biotinylated oligonucleotides (0.5 pmol) were bound to the plate and then pre-incubated protein-primary antibody mixes (in 2/1 Ab/protein molar ratio) were added. Secondary HRP-labeled antibody was incubated on ELISA plate for 30 min, washed and then TMB substrate was added. Absorbance was measured at 370 nm on Synergy H1 (BioTek) and evaluated in GraphPad Prism using hyperbolic or Hill equation fittings. All wash and incubation steps were done in the presence of 2 mM $MgCl_2$ in 1× PBS. Kd were obtained from at least three independent measurements. Details of the procedures are described in [13].

EMSA in agarose gels

scDNAs (200 ng pBSK, pPGM1, pBA50, pPA50) were preincubated in triplex-forming buffer at 37°C for 30 min. scDNAs were mixed with p53 proteins in p53 tetramer/DNA molar ratios 0.25–5 and incubated in binding buffer (5 mM Tris-HCl, pH 8, 1 mM $MgCl_2$, 0.01% Triton X-100 and 50 mM KCl) for 30 min either on ice or 25°C to reach equilibrium. Samples were loaded onto a 1% agarose gel containing 0.33× Tris-borate-EDTA (TBE) buffer. After 5 h electrophoresis (at 4–6 V/cm^2) agarose gels were stained with ethidium bromide (EtBr) and photographed. Intensities of bands of free DNA substrates were quantified using ImageQuant software. Graphs show the evaluation of p53-DNA binding as the dependence of % of bound DNA on the amount of p53 proteins (expressed by molar ratio p53/DNA), more details in [29]. Mean values of three independent experiments were plotted in the graph.

Immunoprecipitation assay

The DO1-p53-DNA complexes were prepared by mixing the DO1 antibody (400 ng) with the purified protein (50 ng) in binding buffer followed by 20 min incubation on ice. Then, 200 ng of scDNA (preincubated in triplex-forming buffer) and the same amount of linDNA (pBSK/*Sma*I) were mixed with the given complexes and incubated in the binding buffer for 30 min on ice. Magnetic beads (12 μl of suspension per sample) coated with protein G (MBG, Dynal/Invitrogen) were added to DO1-p53-DNA complexes after washing in binding buffer and incubated with the beads for 30 min at 10°C. Finally, after washing in binding buffer with increased salt concentration (1× 50 mM, 2× 50–600 mM, 1× 50 mM), DNA was released from the beads by heating at 65°C in 15 μl of 1.0% SDS for 5 min and analyzed by agarose gel electrophoresis, more details in [29]. Intensities of bands of bound DNA substrates were quantified using ImageQuant software. Graphs show the evaluation of p53-DNA binding as the dependence of % of bound DNA on the concentration of KCl. Mean values of three independent experiments were plotted in the graph.

Human cell lines, transfections and luciferase assays

Human breast adenocarcinoma MCF7 (HTB-22, ATCC), human non-small cell lung carcinoma line H1299 (NCI-H1299, ATCC) and H1299-wtp53 cells (Tet-On system, [36]) were grown in DMEM medium supplemented with 5% FBS and penicillin/streptomycin (Gibco). All cultures were incubated at 37°C with 5% CO_2. The luciferase reporter constructs (S1 Table) containing CON and/or $(dA)_{50}$ or $(dA)_{20}$ sequences were used for luciferase assay as described in [29]. pRL-SV40 was used as a transfection efficiency control. 200 ng of reporter construct was transfected in triplicates. Luciferase activity was measured in a plate reader luminometer IMMUNOTECH LMT01 (Beckmann) with Dual Luciferase Assay System (Promega). For each construct, relative luciferase activity is defined as the mean value of the Firefly luciferase/ Renilla luciferase activity ratios obtained from at least three independent experiments.

RT-qPCR

Total RNA was isolated using NucleoSpin RNA II (Macherey-Nagel) and 2 μg of RNA was subsequently reverse transcribed into cDNA by applying High Capacity RT kit (Applied Biosystems). qPCR was performed using EvaGreen (Solis Biodyne) fluorescent dye in the standard program (15 min 95˚C; 15 s 95˚C, 30 s 60˚C, 20 s 72˚C, 10 s 74˚C; 50 cycles) running in Rotor-Gene 6000 (Corbett Research). RT-qPCR reactions for each sample were measured in triplicates. GAPDH was used as reference gene. Absolute quantification was performed using standard curve method. Data were then normalized to GAPDH. The housekeeping genes (HPRT1, GAPDH) were used as endogenous controls. Relative quantification of transcript levels with respect to the calibrator (H1299 with empty vector, MCF7 siRNA control, MCF7) was done based on $2^{-\Delta\Delta CT}$ algorithm. All reactions were carried out in biological triplicates. The primer sequences used are listed in S1 Table.

Immunoblotting

H1299 and Hwtp53 (expressing wtp53, induced with 1 μg/ml tetracycline for 24 hours) cells were harvested from 10 cm plates and lysed with 1× PLB (Promega), followed by the sonication of cells (Bandelin Sonopuls). Samples (100 μg of total protein) were analyzed on 12.5% SDS-PAGE gels and proteins were detected by the following primary antibodies: DO1 (anti-p53, kindly provided by B. Vojtesek), anti-CDKN1A (Millipore), anti-β-Actin (Sigma), anti-BAX (Sigma), anti-NAT10 (ThermoScientific).

Chromatin immunoprecipitation

Human breast adenocarcinoma MCF7 treated for 4 hours with nutlin-3 (5 μM) or doxorubicine (1 μM) were subjected to chromatin immunoprecipitation (ChIP) assays as previously described [29] with the following modifications: the cell sonication was limited to 4 kJ (Bandelin Sonopuls). Purified monoclonal DO1 antibody and IgG (negative control) were incubated overnight with diluted chromatin and immunoprecipitations were performed with protein G-magnetic beads (Invitrogen). The PCR was performed using the primers targeting expected p53 binding site (S1 Table). In other type of ChIP experiment was performed with H1299 cells transfected with plasmids pGL3-PGM1 and pGL3-BA50 (2 μg) and p53 expression vector (pCDNA3.1; 1 μg), after 16 hours cells were subjected to chromatin immunoprecipitation (ChIP) assays. The PCR was performed using the primers targeting expected p53 binding site in pGL3 vector or native promoter sequence see in S1 Table. For quantitative analysis, PCR was carried out for 25 or 30 cycles.

In-silico analysis of promoter regions

Human regulatory sequences were obtained using Table Browser [37] and saved as a FASTA-formatted file of -5000bp to +2000bp regions around each RefSeq TSS. The CON binding sites were identified as closely (<21bp) located pairs of sequence motifs with a maximum of 1 mismatch. The set of identified p53CON sites was expanded to include all full-length grade 3–5 sites identified by p53retriever R/Bioconductor package [38], which largely overlapped the original set. The identification of potential triplex-forming sequences was carried out using the R/bioconductor program triplex-1.8.0 [19], using the default scoring scheme of the software tested in our previous work on human sequences [39]. To check for possible common functions of the identified proteins, we performed a network enrichment analysis using the STRING database tool [40] and gProfiler [41].

In-silico candidate gene transcription screening

Candidate gene transcription was checked in publicly available microarray and sequencing datasets from experiments involving p53-transformed cells originally lacking active p53 or experiments were p53 was activated by nutlin-3, 5-fluoruracil or doxorubicin (SRP043273, SRP022871, E-GEOD-30753, E-GEOD-50650, E-GEOD-8660, E-MEXP-2556 [42]). We obtained expression data from tables available from the iRAP pipeline [43], deposited by authors to Array Express [44] or calculated from the available data using the ArrayExpress R/ Bioconductor package [45]. Raw expression values were normalized relative to GAPDH housekeeping gene and averaged, where replicates were available.

Atomic Force Microscopy (AFM)

AFM measurements were carried out on MultiMode 8 system (Bruker) with NanoScope 8.15 software or on a custom-built AFM system [46]. 50 A silicon nitride MSCT probe, cantilever F (k = 0.5 N/m, Bruker, Santa Barbara, CA, USA), was used with a free amplitude between 1 and 2 nm (amplitude set point between 0.8 and 1.5 nm, 80–90% of the free amplitude). Plasmids were incubated in binding buffer at 37˚C for at least 30 min. For p53-DNA complex images, plasmids were mixed with p53 proteins in p53 tetramer/DNA molar ratio 5/1 and incubated on ice for 20 min. Sample containing 2 ng of plasmid DNA was diluted in 4 mM HEPES pH 7.6, 5 mM $MgCl_2$, 5 mM KCl buffer and placed on freshly cleaved mica V4 surface, incubated for 2 min, washed with distilled water and dried with a stream of compressed air.

Results

Full length p53 binding to intermolecular T.A.*T* triplex is comparable with CTG hairpin non-B DNA structure recognition

Wild type p53 protein is well-known as a non-B DNA structure binder but its interaction with triplex DNA has not been studied yet. We examined p53 binding to pyrimidine type of triplex DNA formed by homoadenine and homothymine oligonucleotides. Intermolecular T.A.*T* triplex was formed in neutral pH in the presence of Mg^{2+} ions [25]. Binding of full-length wild type p53 (p53, Fig 1A) to T.A.*T* triplex was examined by EMSA in the presence of Mg^{2+} ions. Increasing amounts of p53 (50–500 ng, Fig 1A) were bound to 50 bp long random sequence (NON, lanes 2–5), p53 consensus sequence (CON, lanes 7–10) and T.A.*T* triplex (TAT, lanes 12–15). We observed small differences in p53 binding to T.A.*T* triplex (Fig 1A, TAT, lanes 12–15) and to CON (lanes 7–10). In comparison with $CTG_{hairpin}$ (Fig 1B, lanes 7–10) and $TA_{hairpin}$ (Fig 1B, lanes 12–15), the T.A.*T* triplex (Fig 1B, lanes 2–5) was bound by p53 stronger. Considerably weak binding was observed to NON (Fig 1A, lanes 2–5). Detailed titration of p53 protein to T.A.*T* triplex and CON substrates (S1 Fig) mapped the differences between recognition of both substrates.

To better characterize the differences in p53 binding to T.A.*T* triplex in comparison with CON and $CTG_{hairpin}$, we employed an enzyme-linked immunosorbent assay (ELISA) with a set of biotinylated target oligonucleotides CON, TAT and $CTG_{hairpin}$ as recently described for

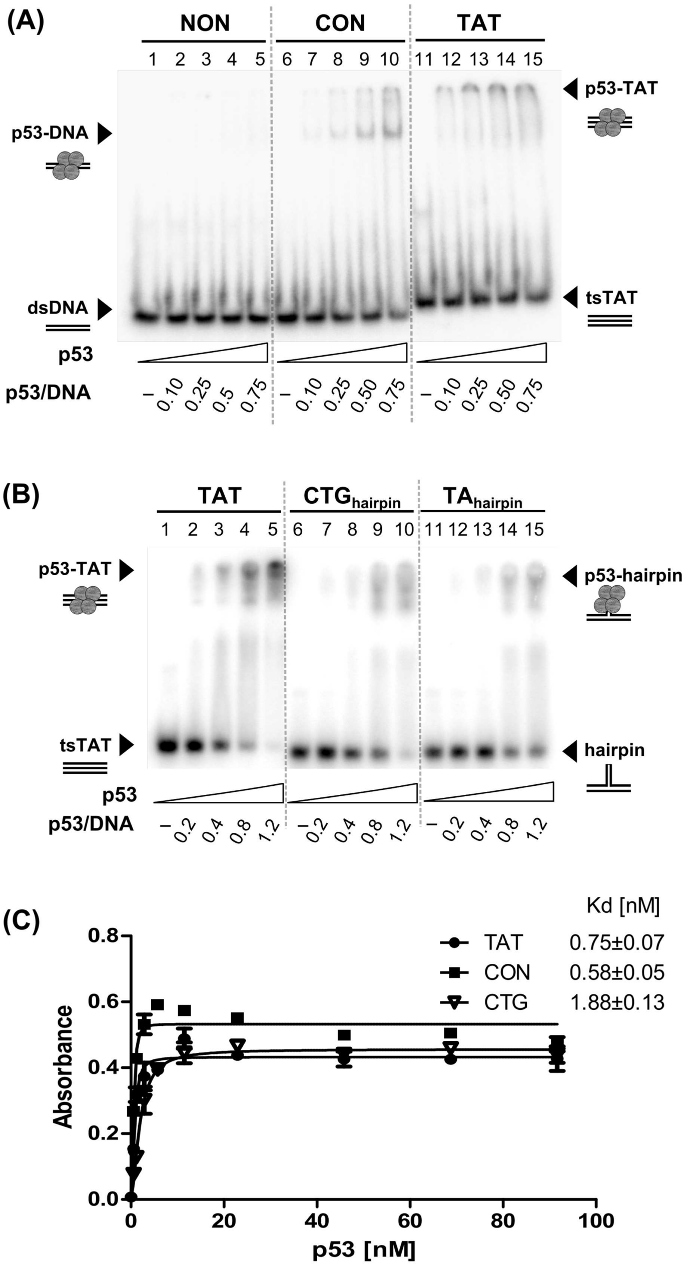


Fig 1. Full length p53 binds strongly to T.A.*T* triplex DNA. (A) Full length p53 was incubated with 1 pmol of ^{32}P-labeled 50-mer oligonucleotides: nonspecific dsDNA (NON, lanes 1–5), p53 specific dsDNA with CON (CON, lanes 6–10) and $(dT)_{50}.(dA)_{50}.(dT)_{50}$ triplex (T.A.*T* triplex, lanes 11–15) in presence of 50 ng pBSK/*Sma*I. Molar ratios of p53 tetramer/DNA ranged between 0.1 and 0.75. The samples were loaded onto 5% 0.5 × TBM (2 mM $MgCl_2$) polyacrylamide gel and electrophoresis was performed for 0.45 h. **(B)** Full length p53 was incubated with 1 pmol of ^{32}P-labeled $(dT)_{50}.(dA)_{50}.(dT)_{50}$ triplex (T.A.*T* triplex, lanes 1–5), CTG hairpin (lanes 6–10) and TA hairpin (lanes 11–15) oligonucleotides in presence of 50 ng pBSK/*Sma*I. Molar ratios of p53 tetramer/DNA ranged between 0.2 and 1.2. The samples were loaded onto 5% 0.5 × TBM (2 mM $MgCl_2$) polyacrylamide gel and electrophoresis was performed for 0.45 h. **(C)** p53 binding to biotinylated oligonucleotides by ELISA. p53 binding curves for the TAT, CON and CTG oligonucleotides are shown, and the dissociation constants (Kd) are indicated.

doi:10.1371/journal.pone.0167439.g001

p53-quadruplex DNA binding [13]. Incubation of the immobilized target oligonucleotides with a range of p53 protein (0.1–90 nM) was followed by quantitation using DO1 antibody. Using this system, we demonstrated that p53 binds to T.A.*T* triplex with higher affinity (Kd = 0.75 ± 0.07 nM), in comparison with $CTG_{hairpin}$ (Kd = 1.88 ± 0.13 nM) (Fig 1C). But as expected, CON (Kd = 0.58 ± 0.05 nM) was the best substrate.

Role of core and C-terminal DNA binding domains for p53 T.A.*T* triplex recognition

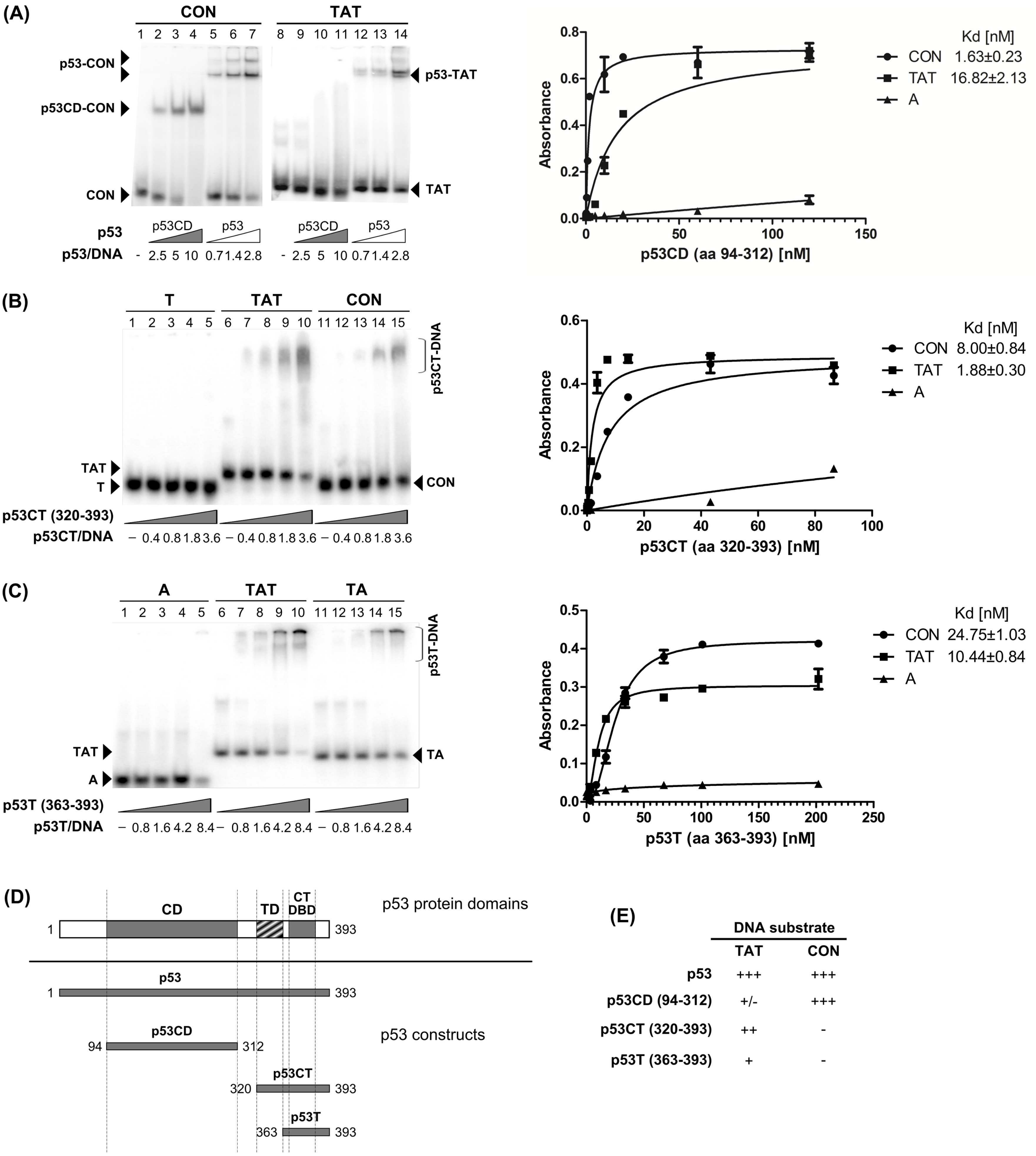
To examine the roles of both DNA binding domains in p53 T.A.*T* triplex recognition we analyzed the interaction of isolated p53 core domain (p53CD, aa 94–312, Fig 2A), C-terminal segment of p53 (p53CT aa 320–393; containing p53CTDBD and tetramerization domains, Fig 2B) and fragment of the last 30 aa of p53 (p53T, aa 363–393, Fig 2C) [10, 29]. At first, we compared binding of p53CD (Fig 2A, lanes 9–11) and full length p53 (Fig 2A, lanes 12–14) to TAT. An unchanged amount of proteins was used for p53CD and p53 binding to CON (Fig 2A, lanes 2–7). In contrast to p53, p53CD was unable to form a stable complex with T.A.*T* triplex.

Binding of C-terminal p53 fragments p53CT (aa 320–393, Fig 2B) and p53T (aa 363–393, Fig 2C) to T.A.*T* triplex was compared with proteins binding to other forms of DNA (ssDNA, dsDNA). We observed that binding of both p53CT and p53T to T.A.*T* triplex DNA was stronger than to the used dsDNA or ssDNA substrates. To better characterize differences in affinities of isolated DNA binding domains to T.A.*T* triplex, we used ELISA with all p53 constructs (p53CD, p53CT and p53T, Fig 2A–2D) followed by quantitation using a specific antibody as was recently described for p53-telomeric quadruplex DNA-binding [13]. With this system, we demonstrated that construct with CTDBD and tetramerisation domain, p53CT (Fig 2B) binds to T.A.*T* triplex with nanomolar affinity (Kd = 1.88 ± 0.30 nM). p53T construct with CTDBD and lacking the tetramerization domain recognized TAT with lower affinity (Kd = 10.44 ± 0.84 nM) than p53CT which is still better than for dsDNA or ssDNA (Fig 2C). And, the lowest affinity for TAT triplex was observed for p53CD (Kd = 16.82 ± 2.13 nM). The results of binding studies are summarized on Fig 2E. Our results showed that the C-terminal DNA binding domain with the tetramerization domain is crucial for TAT triplex high affinity binding.

We confirmed that the C-terminal DNA binding domain is necessary for T.A.*T* triplex recognition by full-length protein with monoclonal antibodies targeting N- and C- terminus (S2A Fig). CTDBD mapping antibody inhibition of p53-non-B DNA complex was previously shown for $CTG_{hairpins}$ and stem-loop structures [33]. DO1, monoclonal antibody targeting aa 20–25 on N-terminus, supershifted both p53-CON and p53-TAT complexes (S2B Fig, lanes 3,8). In contrast to DO1, PAb421 antibody (mapping CTDBD, aa 371–380) induced a partial inhibition of p53 binding to TAT triplex (S2B Fig, lane 9) as opposed to supershifting of p53-CON (S2B Fig, lane 4). ICA9, mapping aa 388–393 on extreme C-terminus, supershifted both p53-CON and p53-TAT complexes (S2B Fig).

Binding of p53 to triplex forming sequence in supercoiled DNA *in vitro*

Intramolecular T.A.*T* triplex (H-DNA) formation in the presence of Mg^{2+} ions in supercoiled plasmids containing homoadenine-homothymine blocks has been described for several vectors [35, 47]. We prepared constructs based on the pBSK vector in variants with and without p53 specific sequence (CON), triplex-forming sequence (TFS, $(dA)_{50}.(dT)_{50}$) and AT-rich cruciform-forming sequence $d(AT)_{34}$ (more details in S1 Table). Formation of non-B DNA structures in different superhelical plasmids was checked by several techniques (S3 Fig): S1 nuclease treatment, OsO_4-bipy modification detected by specific antibody against OsO_4-bipy-DNA adducts [48] (S3F Fig) and OsO_4-bipy modification on the sequencing level [47, 49] (S3G Fig).



	DNA substrate	
	TAT	CON
p53	+++	+++
p53CD (94-312)	+/-	+++
p53CT (320-393)	++	-
p53T (363-393)	+	-

Fig 2. Binding of p53CD and C-terminal p53 fragments to T.A.*T* triplex. (A) p53 Core domain (p53CD, aa 94–312) and full length p53 were bound to CON, (lanes 1–7) and triplex (TAT, lanes 8–14) in p53 tetramer/DNA molar ratios 0.7–10 in presence of 10 ng competitor DNA. Graph of p53CD (aa 94–312) binding to biotinylated oligonucleotides by ELISA. p53CD binding curves for the TAT, CON and A oligonucleotides are shown, and the dissociation constants (Kd) are indicated. **(B)** C-terminal part of p53 (p53CT, aa 320–393) was incubated with $(dT)_{50}$ (T, lanes 1–5), triplex $(dT)_{50}.(dA)_{50}.(dT)_{50}$ (TAT, lanes 6–10) and CON (lanes 11–15) in p53CT tetramer/DNA molar ratios 0.4–3.6. Graph p53CT (aa 320–393) binding to biotinylated oligonucleotides by ELISA. p53CT binding curves for the TAT, CON and A oligonucleotides are shown, and the dissociation constants

(Kd) are indicated. **(C)** C-terminal part of p53 (p53T, aa 363–393) was incubated with $(dA)_{50}$ (A, lanes 1–5), triplex $(dT)_{50}.(dA)_{50}.(dT)_{50}$ (TAT, lanes 6–10) and double-stranded TA (lanes 11–15) in p53CT tetramer/DNA molar ratios 0.8–8.4. Graph of p53T (aa 363–393) binding to biotinylated oligonucleotides by ELISA. p53T binding curves for the TAT, CON and A oligonucleotides are shown, and the dissociation constants (Kd) are indicated **(D)** Scheme showing p53 domains and p53 protein constructs used in this work. **(E)** Relative binding properties of p53 protein constructs to TAT triplex and CON oligonucleotides.

doi:10.1371/journal.pone.0167439.g002

The H-DNAs formed in plasmids pBA50 and pA69 were also visualized by AFM (Fig 3A and S5 Fig).

At first, we compared p53 binding to scDNA capable of H-DNA formation at native superhelical density pBA50 and pPA50 with other plasmids pBSK and pPGM1 by EMSA (Fig 3B). Differences in p53 recognition of scDNA with and without TFS or CON are measurable by number and intensity of retarded bands (compare lanes 3, 8, 13 and 18, Fig 3B) and were evaluated by densitometry of the band corresponding to free (protein-unbound) DNA. The fraction of DNA bound by the protein was calculated and plotted in the graphs shown in Fig 3B (average of at least 3 independent experiments). Both plasmids pPGM1 (with CON, lanes 7–10) and pBA50 (with TFS and H-DNA potential, lanes 12–15) were more strongly bound by p53 than pBSK (Fig 3B, lanes 2–5), similarly to pA69 (with H-DNA potential) versus pUC19 (S4 Fig). The best substrate for p53 was pPA50, plasmid with both motifs CON and TFS (Fig 3B, lanes 16–20).

Furthermore, we applied a competition immunoprecipitation assay and compared binding of p53 to scDNA with and without TFS and CON in the presence of competitor DNA (pBSK/*Sma*I). Increasing salt concentration (50–600 mM KCl, [50]) was applied to detect the difference in stabilities of p53-scDNA complexes containing CON and TFS (Fig 3C). We observed an increase in stability of p53-scDNA binding in the presence of TFS and in agreement with other results, more so in the case of CON (Fig 3B). Due to stability of p53-scDNA complex we were able to perform AFM visualization of p53 bound to scDNA with triplex-forming sequence $(dA)_{69}.(dT)_{69}$ is depicted in Fig 3A and S5 Fig.

To probe differences in relative p53 binding affinity to scDNA with/without TFS and CON we used a competition assay proposed previously [30]. Binding of the p53 protein to CON fragment yielded a well resolved retarded band p53-CON (Fig 3D, lane 2). The intensity of this band was affected by the additions of tested scDNAs, which represented the competitors. Decrease of the p53-CON band intensity relative to the intensity detected in the absence of the competitors reflected the relative affinity of p53 for a given competitor, bar graph represents results from three independent experiments. We observed that pBA50 (T.A.*T*, H-DNA) was a comparable competitor to all plasmids with CON (pPGM1, pPA50 and pPAT34). The control vector pBSK together with pBAT34 (X, cruciform DNA) were the worst competitors.

In-silico screening of human regulatory sequences for co-occurrence of CON binding sites and potential T.A.T triplex-forming sequences

To investigate the possible significance of p53 binding of T.A.*T* triplex-forming sequences for transcription regulation we carried out a series of *in-silico* investigations. Within the context of p53 transcription factor functions involving CON recognition, we looked for T.A.*T* triplex-forming and CON sequence co-occurrence in the human genome to predict new class of p53 target genes. We analyzed the -5000/+2000 bp neighborhoods of 42106 RefSeq gene transcripts (promoters). Of these, 19373 promoters were found to contain at least one CON sequence when 1 mismatch was allowed. T.A.*T* triplex-forming sequences with a prevailing poly(A) or poly(T) run with score> = 18 were found in 376 sequences. Because of the asymmetry in occurrence of these two patterns we decided to screen the promoters primarily on the

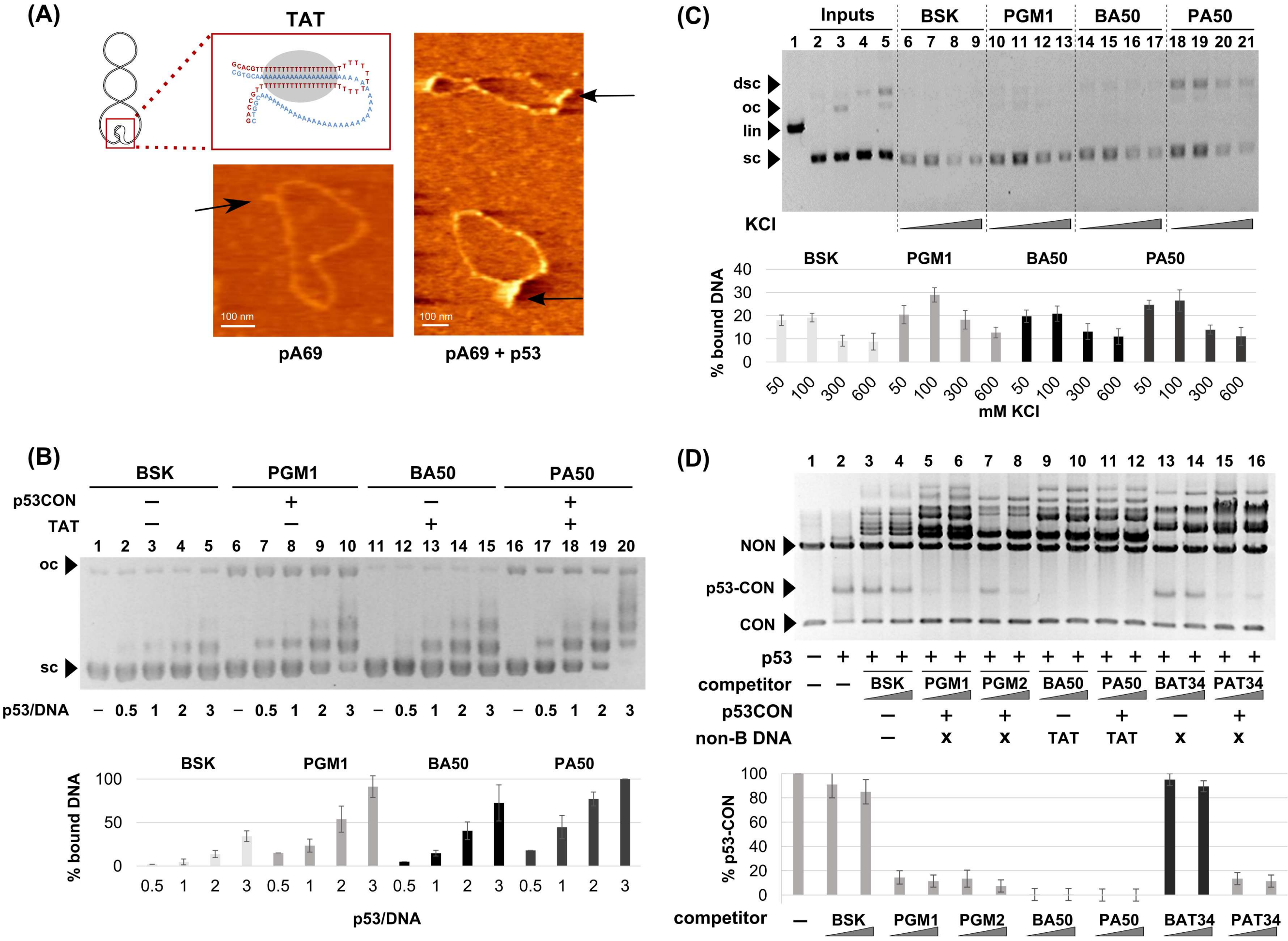


Fig 3. Binding of p53 to supercoiled DNA bearing homoadenine-homothymine triplex forming sequences. (A) Scheme of intramolecular T.A.*T* triplex in scDNA. AFM image of sc pA69 plasmid adsorbed on mica surface in the presence of 2 mM $MgCl_2$ and complex of pA69 with p53. **(B)** Comparison of p53 binding to scDNA with and without triplex forming sequence $(dA)_{50}.(dT)_{50}$ by EMSA. Binding of p53 protein to pBSK, pPGM1, pBA50 and pPA50 detected by EMSA in agarose gel. p53 protein was bound to scDNA (pBSK, 200 ng, lanes 1–5), scDNA with CON (scPGM1, 200 ng, 6–10), scDNA with $(dA)_{50}.(dT)_{50}$ (scBA50, 200 ng, 11–15) and scDNA with both CON and $(dA)_{50}.(dT)_{50}$ (scPA50, 200 ng, 16–20) in p53/DNA molar ratios 1–3 at 4˚C, EMSA was performed at 4˚C. Graph represents the dependence of percents of bound DNA on the amount of p53 proteins calculated from three experiments. **(C)** Interaction of p53 with scDNA (BSK, PGM1, BA50 and PA50) in presence of pBSK/SmaI (linear competitor, lin) by immunoprecipitation on MBG. Agarose gel electrophoresis of DNA recovered from MBG after incubation of DO1-wtp53-DNA complex at the beads to 50, 100, 300 or 600 mM KCl for 30 min at 10˚C followed by the SDS treatment. DNA inputs of scDNA BSK (lane 2), PGM1 (lane 3), BA50 (lane 4), PA50 (lane 5), linBSK (lane 1). Arrows indicate precipitated supercoiled (sc), open circular (oc), linear (lin) and supercoiled dimers (dimer sc). Mean values of bound DNA from three independent experiments were plotted in the graph. Graph represents the dependence of percents of bound DNA on the concentration of KCl in washing buffer calculated from three experiments. **(D)** Competition assay of p53 binding to CON and non-B-DNA structures in scDNA plasmids. First, full length p53 (60 ng) was incubated with 200 ng PGM1/*Pvu*II fragments (short fragment with CON sequence (CON, 474 bp) and long fragment as linear nonspecific competitor (NON, 2513 bp) for 20 min on ice to form p53-CON complexes. Subsequently, 200 or 300 ng of different scDNA plasmid competitors were added and incubation was prolonged to 40 min. Plasmids forming triplex T.A.*T* were marked by TAT, plasmids forming cruciform by X. Graph represents the dependence of percents of bound DNA on the amount of used competitor scDNAs calculated from three experiments.

doi:10.1371/journal.pone.0167439.g003

predicted length of the T.A.*T* triplex. There were 43 promoters of candidate p53 target genes with at least one CON and a T.A.*T* triplex with a poly(A/T) run longer than 40 bp. S2 Table shows locations, common gene abbreviations and binding site data for these promoters. Interestingly, *in-silico* analysis shows that most CONs are downstream of the triplex (Fig 4).

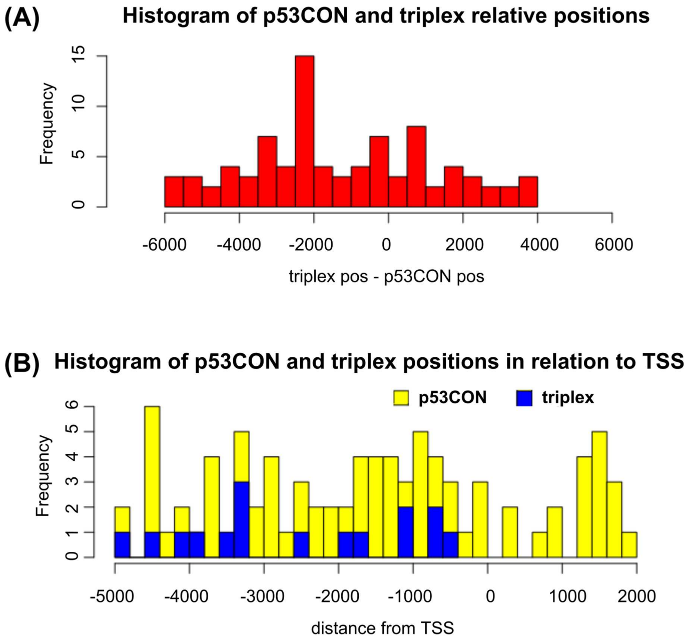


Fig 4. T.A.T triplex and p53CON positions in promoters of the 43 analyzed human genes. (A) Relative distance between each p53CON and the corresponding T.A.*T* triplex. Most p53CONs are 2000-2500bp downstream of the triplex. Second peak corresponds to T.A.*T* triplex positioned in front CON. **(B)** Absolute positions of p53CONs (yellow) and T.A.*T* triplex-forming sequences (blue). TSS is positioned at 0.

doi:10.1371/journal.pone.0167439.g004

STRING-db functional association tool shows 16 of the 43 highest-scoring genes/proteins found in the screening, together with p53 and 10 most-related proteins from STRING-db, organized into a network by common properties and interactions (S6 Fig). 16 proteins from our study that are also part of well-connected networks are: ABCG5, PIK3R4, INSR, MIB1, MAPK9, TGIF1, STAG2, NFAT5, MAK16, DDX54, NAT10, BMS1, PSMB2, PEX12, MCC and MCCC1 shown in blue (S6 Fig). The common functional theme for the proteins clustered by STRING as suggested by gProfiler GO term enrichment analysis is “regulation of signal transduction” (P-value = 2.52e-04).

Triplex forming sequence and DNA topology influence p53 transactivation

To analyze whether the triplex-forming sequence $(dA)_{50}$ has any effect on p53-driven transcription we performed luciferase reporter assays using reporter vectors in variants with and without TFS $(dA)_{50}$, $(dA)_{20}$ too short for triplex formation and p53 specific sequence CON (Fig 5A). Luciferase assay was performed in H1299 cells with transfected pCDNAp53 effector and related to transfected pCDNA vector only (Fig 5B) with linear and supercoiled reporter vectors and in p53 inducible H1299wtp53 cell line (Tet-on system) with sc reporters after p53 induction and related to no induced stage (Fig 5C). Only supercoiled reporters could form non-B DNA structures, in our case H-DNA (Fig 5A, 5B and 5C; B50, P50, TAT) or cruciform (Fig 5A, 5B and 5C; P1, P20, cruciform-X). As expected p53 expression resulted in stronger activation of all vectors containing CON (P1, P20, P50) in comparison with vectors missing CON (BSK and B50). As for P20, with an insert not yet suitable for triplex formation [35], the activation was comparable to the original reporter P1. Interestingly, activation of P50, for intramolecular triplex formation already satisfactory reporter occurring when the reporter was

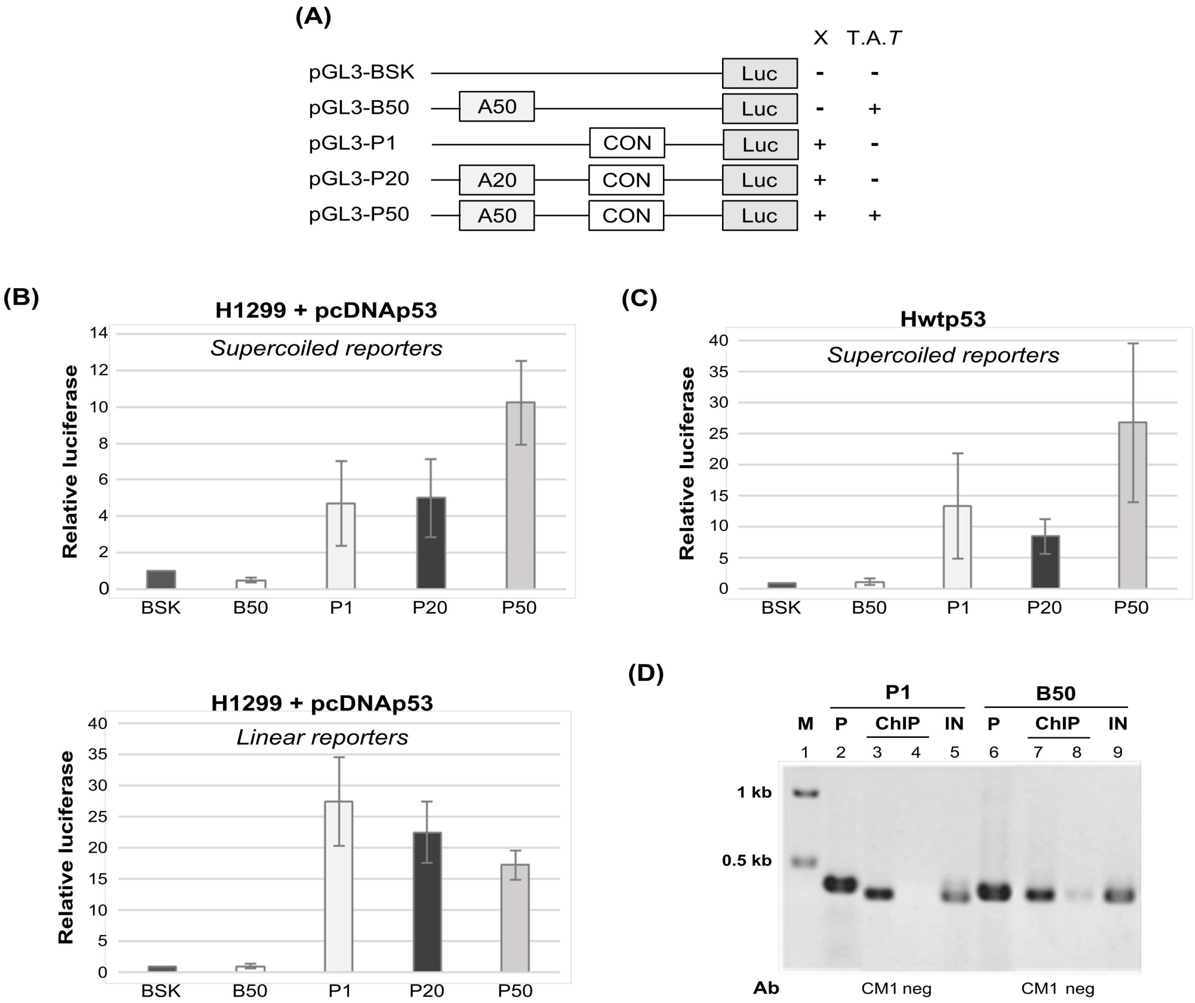


Fig 5. Influence of T.A.*T* triplex forming sequence on p53-driven activation of CON containing reporter vector in scDNA and lin DNA. (A) Scheme of reporter plasmid constructs used in luciferase reporter assay and non-B DNAs formation under supercoiled stress (CF- cruciform, TAT-triplex). **(B-C)** H1299 cells were transiently transfected with plasmids expressing the p53 (pCDNA3.1-p53) or pCDNA3.1 vector alone (CMV) together with reporter: the supercoiled or linear reporter plasmids (BSK, P1, P20, P50, B50) expressing the firefly luciferase gene and a reference plasmid with the renilla gene under control of the SV40 promoter. Luciferase activity was analyzed 16 hours after transfection and signal was normalized on renilla signal. Transfections were carried out in triplicates at least at three independent times and standard deviations are indicated. **(B)** p53 activation of supercoiled reporters. Luciferase activity was normalized on control with vector alone. Only B50 and P50 reporters were able to form triplexes. p53 activation of linear reporter as described above, none of used reporters was able to form triplexes. **(C)** p53 activation of supercoiled reporter plasmids in H1299-wtp53 cells (Tet-on promoter). Luciferase signal after p53 induction was normalized on control without p53 induction. Only B50 and P50 reporters were able to form triplexes. **(D)** Interaction of full length p53 with CON (P1) and triplex T.A.*T* (B50) in scDNA plasmids by ChIP *in vivo*. Plasmids BA50 or PGM1 (2 μg) were transfected into H1299 cells together with vector pCDNA3.1-wtp53 (0.1 μg). ChIP was performed with CM1 antibody. Results of PCR analyses of immunoprecipitated DNA were detected on a 1.5% agarose gel in 1× TAE buffer. PCR samples on the gel are: marker (lane 1), plasmid PGM1 (P1, lane 2) and BA50 (lane 6); 1/20 of DNA input (lanes 5 and 9 marked as IN); IP with IgG (negative control) (lanes 4 and 8); IP with CM1 Ab (lanes 3 and 7).

doi:10.1371/journal.pone.0167439.g005

supercoiled, was significantly stronger than analogous reporter containing only CON (P1) (Fig 5B and 5C). For linear reporter P50 such effect was not observed (Fig 5B). In the case of B50, a repression was observed with sc form of reporter (Fig 5B and 5C). In summary, triplex-forming sequence $(dA)_{50}$ enhances p53-driven transcription from supercoiled reporter containing p53 specific sequence CON.

To confirm *in vivo* p53 binding to $(dA)_{50}$ sequence capable to form H-DNA, supercoiled plasmids B50 (H-DNA potential) and P1 (CON with potential to form DNA cruciform) were transfected to H1299 cells together with effector plasmid pCDNA3.1p53 and a ChIP assay was performed with p53 specific antibody CM1 (Fig 5D lane 3 and 7) and IgG (negative control) (Fig 5D, lane 4 and 8). We observed comparable binding of p53 to B50 (TAT, H-DNA-forming sequence Fig 5D, lane 7) as to P1 (CON, Fig 5D, lane 3).

Together, these data demonstrate that the triplex-forming sequence $(dA)_{50}$ under conditions favorable for the actual H-DNA formation can influence the level of DNA-binding and transactivation of p53 binding sites in promoter regions by p53 *in vivo*.

Analysis of candidate p53 target genes with triplex-forming sequences in promoter region

To better prioritize the candidate p53 target genes identified by the above *in-silico* screening (S2 Table and S6 Fig) we consulted publicly available microarray and sequencing datasets for experiments involving full-length p53, p53CΔ30 and p53S389A transformed cells originally lacking p53 [51, 52] or experiments with endogenous p53 activated by nutlin-3/doxorubicin/5-fluoruracil for gene expression values [42, 53–58], results are summarised in S3 Table. This way we were able to evaluate expression of many of the candidate p53 target genes and also evaluate the influence of p53 C-terminus as shown in S3 Table. Several of the genes selected by the screen showed consistent up-regulation in these conditions (*MCC*, *NFAT5*, *ENOX1*, *ABCG5*) or down-regulation (*MAPK9*, *MAK16*). Interestingly, *NAT10* and *STAG2* belongs to several genes down-regulated after activation of p53 by drug treatment and up-regulated in p53 overexpression in p53 null cells. Several up or down regulated genes (*ABCG5*, *INSR*, *MCC*, *NFAT5* and *NAT10*) were limited to the STRING-db-supported functionally associated group of genes. Intact C-terminus was necessary for strong p53-dependent activation of *MCC*, one of the best candidate p53 target gene, in contrast to well-known target gene *MDM2* (S3 Table).

To validate experimentally our set of candidate genes (S2 Table) as novel p53 target genes, at first we performed their RT-qPCR analysis after p53 transient transfection experiment in p53 null cell line (H1299, Fig 6A, S3 Table). As expected p53 overexpression activated *p21*, *BAX* and several new potential candidate p53 target genes (e.g. *ABCG5*, *INSR*, *MCC*, *NFAT5*; Fig 6A). Next, we checked whether p53 downregulation in MCF7 cells could reduce their expression. Downregulation after p53siRNA treatment was observed for *ABCG5*, *ENOX1*, *INSR*, *MCC*, *NAT10* and *NFAT5* (Fig 6A, S3 Table). In addition, *ABCG5*, *ENOX1*, *INSR*, *MCC*, *NFAT5* together with *p21* and *BAX* were induced in MCF7 cells treated with nutlin-3, a p53-stabilizing agent (Fig 6B). However, activating p53 by actinomycin D did not promote *ENOX1*, *INSR*, *MCC* expression, in contrast to *BAX*, *p21* and *ABCG5* (Fig 6B). For another candidate genes *MAPK9* and *NAT10* we observed down-regulation after p53 activation by actinomycin D drug treatment. Interestingly, after 24 hours tetracycline p53 induction of Hwtp53 cells, we observed activation of NAT10, p21 and BAX on the protein level (Fig 6C). To determine binding of endogenous p53 to triplex forming sequences in selected new potential p53 target gene promoters, we performed ChIP assay for analysis of p53 binding on *MCC*, *NAT10* and *p21* promoters in MCF7 cells (Fig 6D). Using of primers covering TAT triplex we observed p53 binding to *MCC* and *NAT10* promoters also after stabilization of p53 after nutlin-3 and doxorubicin treatment in MCF7 cells (Fig 6D). Taken together, *in silico* analysis of expression data, RT-qPCR and ChIP analysis have shown connection between p53 and new set of potential p53 target genes with triplex forming sequences in promoter regions.

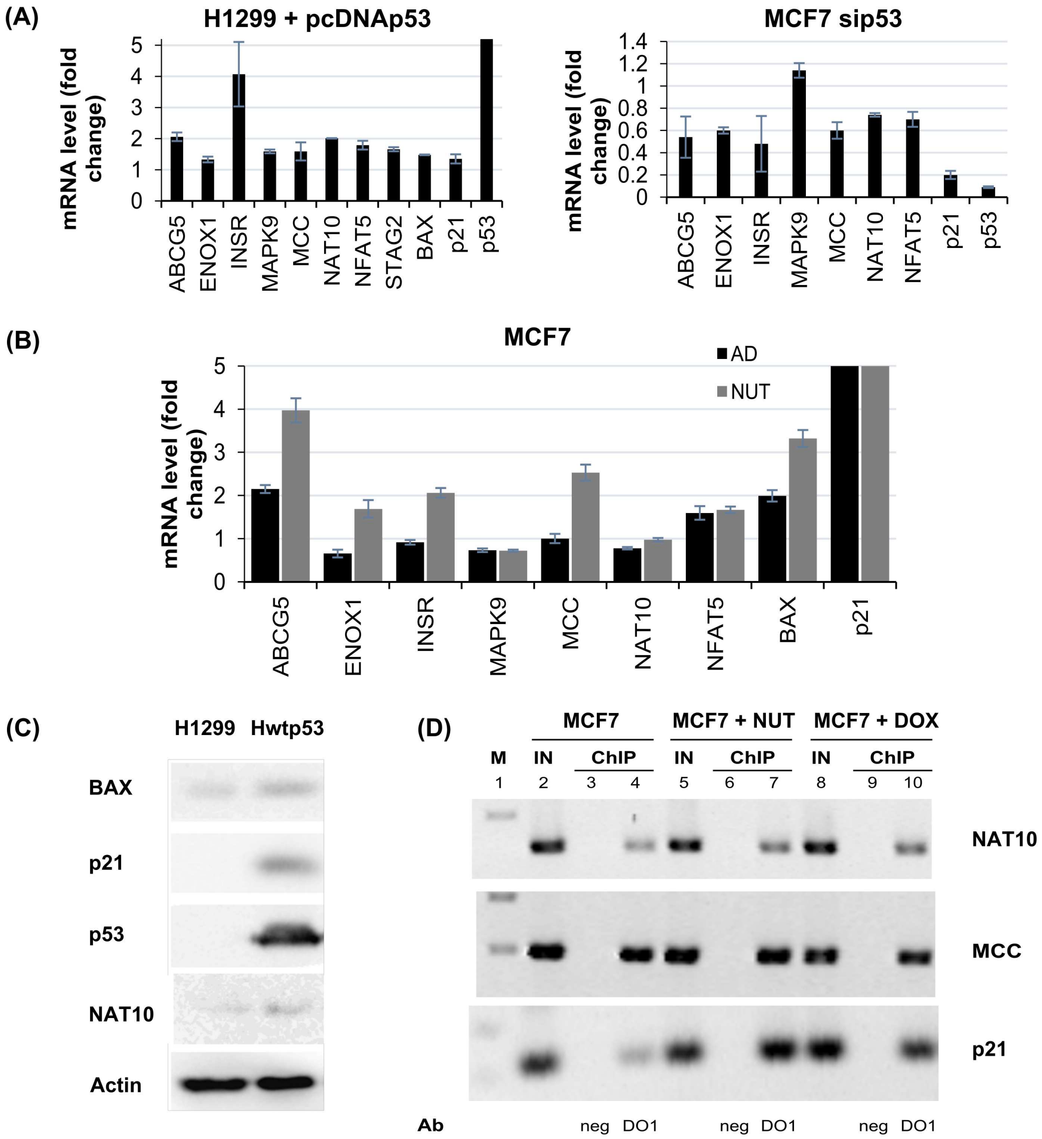


Fig 6. Verification of candidate p53 target genes. (A) RT-qPCR analysis of candidate p53 target genes and BAX, p21 and p53 mRNA levels in i) H1299 cells transfected by pCDNAp53 for 48 hours (left graph); ii) MCF7 cells with downregulation of p53 by siRNA over control siRNA for 48 hours (right graph). **(B)** RT-qPCR analysis of candidate p53 target genes and *BAX*, *p21* mRNA levels in MCF7 cells after nutlin-3 or actinomycin D 12 hours treatment. Gene values were normalized to GAPDH. The values are the average of three independent experiments. **(C)** p53 mediated up-regulation of NAT10 on protein level and activation of BAX and CDKN1A was analyzed in Hwtp53 cells (24 hours induction) vs H1299 without p53 expression. Western blots presenting protein levels of p53, NAT10, CDKN1A and BAX. Actin was used as loading control. **(D)** Chromatin immunoprecipitation showing p53 binding to *MCC* and *NAT10* promoters which contain a TAT triplex motif. DNA fragments from MCF7 cells without and with nutlin-3/ doxorubicin 4 hours treatment were immunoprecipitated using DO1 antibody against p53 (lane 4,7,10), negative control ChIP with IgG (lanes 3,6,9), positive input control (1/15 input for ChIP, lanes 2,5 and 8).

doi:10.1371/journal.pone.0167439.g006

Discussion

Alternative, non-B DNA structures, such as triplex, quadruplex, hairpin and cruciform can be formed by sequences that are widely distributed throughout the human genome [59]. Triplexes and cruciforms are implicated in regulating gene expression and causing genomic instability

[60, 61]. Despite the known fact of tumor suppressor p53 protein importance for maintaining genomic stability, the mechanisms in this protective function are still not well understood.

Regions with the potential to form triplex DNA are generally over-represented in the promoter regions and introns of genes involved in cell signaling as indicated by genome-wide bioinformatics analyses [18, 19, 62]. In our previous bioinformatics study, we showed the prevalence of the T.A.*T* triplex class in the human genome [19]. The present work was a follow-up by focusing on p53 recognition of T.A.*T* triplex-forming sequence $(dA)_{50}.(dT)_{50}$, especially in promoters containing this sequence in close proximity to specific p53 binding sites (CONs).

A number of independent studies have established that p53 recognizes non-B DNA structures including hairpins, stem-loops, cruciforms, mismatches, bulges, G-quadruplexes, three- and four-way junctions [4, 30, 31, 63–66]. For example CTG.CAG trinucleotide repeats were shown to be a novel class of p53-binding sites *in vitro* and *in vivo*, CTG and CAG hairpins were determined as p53 bound non-B DNA structures in that repetitive sequence [33]. To best of our knowledge no study has been published on triplex DNA recognition by wild-type p53 protein. Mutant p53 (R273H) binding to genomic fragment containing mirror repeats with the potential to form intramolecular triplex was shown in an earlier study of ours on identifying natural binding sites in glioblastoma cell line U251 [67].

In the present study, a range of biophysical approaches was used to analyze the interaction of full-length and isolated DNA binding domains of p53 with intermolecular triplex DNA. The T.A.*T* type of triplex was chosen with respect to physiological conditions necessary for triplex formation [35, 47] and for the high frequency of potential triplex-forming sequences in the genome [39]. Both EMSA and ELISA assays demonstrate slightly greater binding affinity of full-length p53 protein to the T.A.*T* triplex than to the $CTG_{hairpin}$ (Fig 1). Binding of full-length p53 to T.A.*T* triplex was weaker than to specific sequence CON. In contrast to p53T and p53CD, the affinity of p53CT for the T.A.*T* triplex was in range of full-length p53. Thus, our data showed that both CTDBD and the tetramerization domain (aa 325–356) are necessary for high affinity p53 binding to the T.A.*T* triplex.

Although binding of DNA by the C-terminus is usually marked as non-specific, CTDBD has a major role in non-B DNA structures recognition (e.g. stem-loop structure, G-quadruplex, CTG and CAG hairpins, [13, 31, 33, 68]) and there is increasing evidence for the importance of intact CTDBD for regulating sequence-specific DNA binding, transactivation and also for the maintaining genomic stability [69, 70]. The C-terminus is marked by the presence of a large number of positively charged amino acid residues and has an inherently disordered character. The CTDBD structure gives intrinsic flexibility and possesses molecular recognition features necessary for the multifunctional nature of this region [70, 71]. The formation of a partially helical structure was observed experimentally after binding of the C-terminus to non-specific DNA (sheared herring sperm DNA, [72]). Laptenko´s recent *in vivo* and *in vitro* study with p53 proteins mutated in CTDBD (mimicking acetylation/phosphorylation) points to several positive roles of intact unmodified CTDBD in regulating sequence specific DNA binding, p53 protein stability, p53 cellular localization and co-factor recruitment [70]. Recently, the relevance of post-translational modifications of the C-terminus in the DNA-binding properties of p53 has been reviewed in [71].

There is no systematic study to date of the role of DNA binding domains in different non-B DNA structures recognition. CTDBD is necessary for recognition of DNA cruciform and stem-loop structures both formed by CON sequences [30, 31], as well as CTG.CAG tracts [33]. In the case of p53 interaction with scDNA, we have shown that at least the dimeric form of CTDBD is essential for highly selective binding [10]. Three-stranded junctions (with and without mismatches) were recognized by full length protein but with lower affinity by p53CΔ30 (containing core domain with the tetramerization domain) as well [4]. On the other hand, the CD and dimerization domain are required for high affinity interaction with insertion/deletion lesions [5]. Our data agree with the majority of studies on p53 interaction with alternative DNA structures, showing the CTDBD and tetramerization domain is responsible for high-selective binding of p53 to non-B DNA structures [4, 9, 12, 30–32, 73].

For the first time we show preferential p53 binding to supercoiled plasmids capable of H-DNA formation by $(dA)_{50}.(dT)_{50}$ sequence. We verified H-DNA formation under superhelical stress under conditions used for p53 binding using several techniques and visualized them by AFM. scDNA pBA50 was somewhat more weakly bound by p53 than scDNA with CON (pPGM1). In competition assay, pBA50 and pPA50 capable of H-DNA formation were better competitors than $pBAT_{34}$ forming AT-rich cruciform and comparable in competition to plasmids with CON (Fig 3D). Supercoiled pPGM1 was shown to form cruciform by CON with stem-loop motif with mismatches and to be more attractive for p53 binding [30, 31]. We suspect that the high affinity of p53 for scDNA capable of forming H-DNA is due to the fact that besides the triple-helical part of the scDNA molecule (Fig 3A), p53 also recognizes single-stranded loops and junctions (Fig 3A) already described as p53 recognition motifs in DNA [31].

Identification of T.A.*T* triplex as a novel p53 binding site recognized by CTDBD raises the question of the physiological significance of such interaction. The nM binding/dissociation constant that we observed for p53 binding to intermolecular T.A.*T* triplex (Fig 1C) shows that this binding is slightly stronger than to $CTG_{hairpin}$ and slightly weaker than to CON observed in this work using ELISA and EMSA (Fig 1) providing evidence for the *in vivo* relevance T.A.*T* triplex p53 binding. The nM range of binding/dissociation constant for p53 sequence-specific interaction has been found by several groups using various techniques e.g. Fersht´s group by FA [30, 31]. For sequence-specific p53 binding, application of competitive fluoresce anisotropy technique has shown Kd values in the range of 10–100 nM. The pM dissociation constant for sequence-specific and insertion/deletion lesion p53 interactions has been reported so far in only one study [5].

We speculate that the T.A.*T* triplex formed by $(dA)_{50}.(dT)_{50}$ tracts may act as a non-B DNA p53 binding site essential for p53 stability, co-factor recruitment and regulating sequence-specific binding mainly in the case of unmodified C-terminus by phosphorylation and acetylation. Binding of p53 to a significant number of sites within the genome depends on the availability of unmodified CTDBD according to a recent report [70]. The C-terminus has been shown to be crucial for the sliding mechanism of p53 recognition of CON by p53CD [74]. p53 binding to multiple non-B binding sites can influence their stability. One suggested scenario is that non-B DNA structures may be targeted by p53, which then binds to and stabilizes or destabilizes such DNA structures to increase gene transcription. Besides its effect on gene transcription, p53-non-B DNA recognition can participate in DNA repair, DNA replication and/or DNA recombination. Genome-wide studies show that p53 binds to many loci in the genome, including sites not associated with transcriptional control [75]. Recently, the prevention of accumulation of DNA damage by p53 binding to subtelomeric regions has been described [76]. Walter et al. showed that p53 induces local distortions in mismatched trinucleotide repeats and suggested that p53 may be involved in the maintenance of CTG.CAG tract stability [12, 30, 33]. In our case we observed a positive effect of T.A.*T* triplex-forming sequence $(dA)_{50}.(dT)_{50}$ on the stability of the p53-scDNA complex and p53 binding to $(dA)_{50}.(dT)_{50}$ in scDNA in cells. For this reason, we hypothesize that p53 interaction with T.A.*T* triplex, primarily by CTDBD, can stabilize p53 protein in both non-B DNA and CON. Additionally, we can discuss the role of the p53-T.A.*T* triplex recognition in the process of DNA repair. It was shown that triplex-forming oligonucleotides are able to activate DNA recombination and

DNA repair in addition to inducing genomic instability [77]. Intact p53 C-terminus is necessary for recognition of damaged DNA and recombination intermediates [2, 3, 7, 8, 63, 78, 79]. Triplex DNA may also elicit genetic instability by a roadblock to DNA replication and transcription elongation [80]. The DNA damage tolerance pathway and p53 regulates DNA replication fork progression according to a recent study [78]. It was shown, that the helical distortions and structural alternations induced by triplex formation may be recognized as "DNA damage" [80, 81]. So far, we can only speculate that p53-T.A.*T* triplex recognition can eliminate DNA damage caused by triplex formation.

Interestingly, the group of proteins specifically recognizing triplex DNA (HMG, helicases, RAD51, RPA [82]) are also known as p53 interaction partners. As large number of p53 interacting proteins also interact with triplex DNA, we reason that p53 triplex recognition has the potential to influence the regulation of genomic stability, DNA repair, DNA replication, DNA recombination and gene expression at different levels.

Using luciferase reporter assay in two different cell systems, we demonstrate that T.A.*T* triplex-forming sequences $(dA)_{50}.(dT)_{50}$ in front of CON, enhanced promoter activation by p53. Interestingly, the reporter vector containing only T.A.*T* triplex-forming sequence $(dA)_{50}.(dT)_{50}$ was repressed by p53 protein. Both these effects suggested that T.A.*T* triplex-forming sequences have the potential to influence transcription in both directions. We assume that positioning of p53 on promoter region facilitates p53 recognition and transcription of genes.

Our *in-silico* analysis with STRING showed that a fraction of promoters containing both CON and a potential T.A.*T* triplex-forming sequence belong to the functional and structural association network of p53. Although p53 has a large association network, repeated experiments with randomly chosen UniProt Ids have shown that the majority of blind tests had networks with less than 10 interactions while we observed 14, before adding the additional 10 best connected proteins. A medium strength enrichment (P-value ~ 0.00025 after correction for multiple testing) was obtained from gProfiler for the most enriched Gene Ontology term: "regulation of signal transduction". Consequently, the *in-silico* experiments did not yield results that would have the power of proof for us. Rather, they should be viewed as a tool to narrow down possible candidates for further studies, such as the RT-qPCR experiments carried out here. Several candidate genes from the narrowed-down list that have been tested by RT-qPCR show increased expression in p53 dependent manner in p53 null cell line. The best candidates are *ABCG5*, *ENOX1*, *INSR*, *MCC*, *NAT10*, *NFAT5* and *MAPK9* (Fig 6). Only *MCC*, *INSR* and *NAT10* association with p53 has been described so far. *MCC* was described as a target gene upregulated by nutlin-3 but not by doxorubicin and its promoter CON sequence was bound by p53 in U2OS cells [83]. *INSR* is described as a target gene upregulated by overexpression of p53 in HCT116 p53-/- cells [83]. Recently, NAT10 was described as a protein regulating p53 activation through its acetylation and also that NAT10 was upregulated under stress conditions in a p53-dependent manner. Thus, NAT10 forms a positive regulation feedback with p53 in response to stress [84].

The tumor suppressor p53 has been studied extensively as a direct transcription regulator of several hundred target genes and it is currently known to indirectly regulate thousands of genes [85]. Detailed promoter analyses of each potential candidate p53 target gene have to be done to validate them as genuine p53 target genes, as well as, to prove the importance of DNA triplex formation for their regulation by p53. So far, *in-silico* analysis of promoters of candidate p53 target genes shows that most CONs are downstream of the triplex and we can only speculate about the possible functions of T.A.*T* triplex-forming sequence as enhancers and this has to be experimentally proven. Recently, p53 recognition of regulatory enhancer elements within the non-coding genome was identified in human fibroblasts [86]. p53 has been shown to regulate the expression of multiple genes over long distances via looping and binding to enhancers

[85]. Originally, we showed that p53 is involved in DNA looping *in vitro* [87]. More experiments with positioning of TAT and CON sequences have to be conducted to confirm this hypothesis.

Genome organization and local DNA structural effects on gene expression are still not sufficiently investigated. Our results show possible concomitant binding modes of p53, where one of them depends on structures that may only be present transiently in the genome. Further studies would provide us with better understanding of the local environment at promoters and new modes of transcriptional regulation.

Conclusions

In summary, we show that p53 protein possessing intact C-terminus exhibits the ability of p53 to bind with high affinity to intermolecular and intramolecular T.A.*T* triplex DNA. Moreover, T.A.*T* triplex influences transcription from a CON containing reporter and p53 T.A.*T* binding was also detected *in vivo* by chromatin immunoprecipitation techniques. *ABCG5*, *ENOX1*, *INSR*, *MAPK9*, *MCC*, *NAT10* and *NFAT5* were associated with p53, as potential novel p53 target genes with T.A.*T* motif in their promoter.

Supporting Information

S1 Fig. Protein analysis and comparison of binding of full length p53 to T.A.*T* triplex and CON. (A) SDS-PAGE analysis of p53 proteins used in the study. The purity and appropriate size of each proteins was analyzed by Coomassie blue staining of 12.5% SDS-PAGE gel. **(B)** Full length p53 was bound to 1 pmol of ^{32}P-labeled 50-mer oligonucleotides represented by p53 nonspecific dsDNA (NON, lanes 1–5), p53 specific dsDNA with CON (CON, lanes 6–10) and triplex $(dT)_{50}.(dA)_{50}.(dT)_{50}$ (TAT, lanes 11–17) in the presence of DNA competitor (linear plasmid pBSK/*Sma*I, 50 ng). The reactions were separated on 4% 0.5× TBM (2 mM $MgCl_2$) polyacrylamide gel (PAGE), 3h. Radioactively labeled DNA was detected by autoradiography. B,C) Full length p53 was bound to 1 pmol of ^{32}P-labeled 50-mer oligonucleotides represented by p53 specific dsDNA with CON (CON, B) and triplex $(dT)_{50}.(dA)_{50}.(dT)_{50}$ (TAT, C) in the presence of DNA competitor (linear plasmid pBSK/*Sma*I, 20 ng). The reactions were separated on 5% 0.5× TBM (2 mM $MgCl_2$) PAGE, 1 h. Radiolabeled DNA was detected by autoradiography.
(TIFF)

S2 Fig. Interaction of CTDBD with T.A.*T* triplex. The effect of C-terminal modifications of p53 protein by Ab on T.A.*T* triplex recognition. (A) Scheme of p53 used in this study, shown as boxes below the map of p53 domains. The evolutionarily conserved domains are indicated: core DNA binding domain (CD; aa ~100–300), tetramerization domain (TD; aa 325–356) and basic C-terminal DNA binding domain (CTDBD; aa 363–382) and location of p53 antibodies PA421, ICA9 and DO1 used in our study. **(B)** Effect of C-terminal modifications of p53 protein by Ab on T.A.*T* recognition. The antibodies (DO1, PAb421 and ICA9; 1.5 μg) were bound to p53 (300 ng) in Ab/p53 molar ratio 2/1 at RT for 15 min. Then 1 pmol of ^{32}P-labeled 50-mer oligonucleotides represented by p53 specific dsDNA with p53CON (CON, lanes 1–5) and triplex $(dT)_{50}.(dA)_{50}.(dT)_{50}$ (TAT, lanes 6–10) were added and mixtures were incubated at 4°C for 20 min. The reactions were separated on 4% 0.5× TBM (2 mM $MgCl_2$) PAGE at 4°C. Radioactively labeled DNA was detected by autoradiography. Mouse monoclonal anti-p53 antibodies (mAb) (DO1 (aa 20–25), Bp53 10.1 (aa 375–379), PAb421 (aa 371–380) and ICA9 (aa 388–393)) and anti-GST Ab (G1160, Sigma) were used.
(TIFF)

S3 Fig. Non-B DNA structures analysis supercoiled plasmid DNA (pBSK, pPGM1, pPGM2, pBA50, pPA50, pBAT34, pA69 and pPAT34) by S1 treatment, OsO4-bipy modification and its combination with S1 treatment. (A,B,D,E) Scheme of non-B DNA structures detection by S1 nuclease treatment described in [30]. scDNAs were treated with S1 nuclease followed by *Sca*I digestion. Detection of two fragments indicates one major non-B DNA structure (cruciform or triplex) formation in the polycloning site in the case of pBA50 (A), pPGM2 (D, lane 12), pBAT34 (E, lane 4), and pAT34 (E, lane 8). But also pPGM1 (D, lane 8), pBA50 (B; E, lane 12), pPA50 (E, lane 16) and pBSK (D, lane 4) were sensitive to S1 nuclease treatment; two pairs of fragments (black lines) were detected, indicating that all plasmids can form non-B DNA structures with unpaired bases. (C) AFM visualization of intramolecular triplex in pBA50, conditions as described in Fig 2. (F) Detection of non-B DNA modified with OsO_4-bipy by dot blot on nitrocellulose membrane with specific antibody against OsO_4-bipy-DNA adduct as described in [48]. pUC19 (vector only) and pA69 were modified by condition described in [48]; (G) Detection of non-B DNA in plasmid DNA pre-incubated in 20 mM TrisHCl pH8, 2mM $MgCl_2$ without/with 100 mM NaCl by OsO_4-bipy modification followed by primer extension analysis of pBSK (1,2), PGM1 (3,4), PGM2 (9,10), pBA50 (11,12) plasmid DNA, conditions described in [47]. Primer extension from T7 primer was used. See S1 File for experimental details.
(TIFF)

S4 Fig. Comparison of p53 binding to scDNA with and without triplex forming sequence $(dA)_{69}$.$(dT)_{69}$ by EMSA. Binding of p53 protein to pUC19 and pA69 detected by EMSA in agarose gel. p53 protein was bound to scDNA (pUC19, 200 ng, lanes 1–5) and scDNA with $(dA)_{69}(dT)_{69}$ (pA69, 200 ng, 6–10) in p53/DNA molar ratios 1–5 at 25°C, EMSA was performed at 4°C.
(TIFF)

S5 Fig. AFM visualization of plasmids containing triplex-forming sequences and their complexes with p53 proteins. (A) AFM image of scBA50 plasmid mounted in the presence of 5 mM $MgCl_2$. Scale bar represents 200 nm. **(B)** Image of pA69 complexes with p53, proteins were incubated with DNA in molar ratio 5/1 in DNA binding buffer and then loaded on mica surface in the presence of 5 mM $MgCl_2$. Scale bar represents 500 nm. **(C)** pA69 plasmid with p53 proteins in 3D projection.
(TIFF)

S6 Fig. STRING-db analysis of the highest-scoring proteins of candidate p53 target genes. The 43 highest-scoring proteins of candidate p53 target genes found in the *in-silico* study (red and blue), together with p53 (yellow) and 10 most-related proteins (grey) from STRING-db, organized into a network by common properties and interactions. The 16 proteins from our study that are also part of well-connected networks are shown in blue. See S1 File for experimental details.
(TIFF)

S1 File. Supplementary Methods.
(DOCX)

S1 Table. Sequences of DNA oligonucleotides, DNA plasmids and primers for ChIP and qRT-PCR, separate file.
(XLSX)

S2 Table. Tabulated positions of identified p53CON and longest T.A.*T* triplex sequences relative to the transcription start site of the given RefSeq transcript. Positions of lower

stringency p53CON sequences with 2 mismatches are shown in parentheses. Genome coordinates refer to human genome sequence hg38 annotation.
(XLSX)

S3 Table. Verification of candidate p53 target genes. *In-silico* candidate gene screening of publicly available microarray and sequencing datasets and summarization of results of verification by RT-qPCR. See S1 File for experimental details.
(XLSX)

Acknowledgments

We are indebted to Drs. C. Midgley, M., K. Walter, K.R. Fox, D. Klein, T. Hupp and Prof. G. Selivanova for the supply of plasmids. Technical assistance by L. Holaňová is acknowledged. Dr. K. O. van der Werf for help with AFM instrumentation. Prof. E. Paleček, Dr. M. Zimmerman, Prof. M. Vorlíčková, Dr. P.Pečinka and Dr. M. Fojta are acknowledged for discussions and recommendations during the project.

Author Contributions

Conceptualization: MB ML VS MLB.

Data curation: ML TM.

Formal analysis: ZB RH ML TM.

Funding acquisition: MB ML TM VS.

Investigation: MB VT RH PB AP MP LN OT KN MLB ZB TM ML MA.

Methodology: MB VT PB ML RH.

Resources: MB VS MLB ML TM.

Supervision: MB.

Validation: ZB RH.

Visualization: MB VT RH AP AK MP LN ML TM.

Writing – original draft: MB ML VT MA MP LN.

Writing – review & editing: MB VT RH PB AK MP LN OT KN MLB VS ZB TM ML MA.

References

1. Vogelstein B, Kinzler KW. p53 function and dysfunction. Cell. 1992; 70(4):523–526. PMID: 1505019
2. el-Deiry WS, Kern SE, Pietenpol JA, Kinzler KW, Vogelstein B. Definition of a consensus binding site for p53. Nat Genet. 1992; 1(1):45–49. doi: 10.1038/ng0492-45 PMID: 1301998
3. Bakalkin G, Selivanova G, Yakovleva T, Kiseleva E, Kashuba E, Magnusson KP, et al. p53 binds single-stranded DNA ends through the C-terminal domain and internal DNA segments via the middle domain. Nucleic Acids Res. 1995; 23(3):362–369. PMID: 7885831
4. Dudenhoffer C, Kurth M, Janus F, Deppert W, Wiesmuller L. Dissociation of the recombination control and the sequence-specific transactivation function of P53. Oncogene. 1999; 18(42):5773–5784. doi: 10.1038/sj.onc.1202964 PMID: 10523858
5. Szak ST, Pietenpol JA. High affinity insertion/deletion lesion binding by p53. Evidence for a role of the p53 central domain. The Journal of biological chemistry. 1999; 274(6):3904–3909. PMID: 9920946

6. Jett SD, Cherny DI, Subramaniam V, Jovin TM. Scanning force microscopy of the complexes of p53 core domain with supercoiled DNA. Journal of molecular biology. 2000; 299(3):585–592. doi: 10.1006/jmbi.2000.3759 PMID: 10835269

7. Zotchev SB, Protopopova M, Selivanova G. p53 C-terminal interaction with DNA ends and gaps has opposing effect on specific DNA binding by the core. Nucleic Acids Res. 2000; 28(20):4005–4012. PMID: 11024181

8. Fojta M, Pivonkova H, Brazdova M, Kovarova L, Palecek E, Pospisilova S, et al. Recognition of DNA modified by antitumor cisplatin by "latent" and "active" protein p53. Biochem Pharmacol. 2003; 65 (8):1305–1316. PMID: 12694871

9. Stros M, Muselikova-Polanska E, Pospisilova S, Strauss F. High-affinity binding of tumor-suppressor protein p53 and HMGB1 to hemicatenated DNA loops. Biochemistry. 2004; 43(22):7215–7225. doi: 10.1021/bi049928k PMID: 15170359

10. Brazdova M, Palecek J, Cherny DI, Billova S, Fojta M, Pecinka P, et al. Role of tumor suppressor p53 domains in selective binding to supercoiled DNA. Nucleic Acids Res. 2002; 30(22):4966–4974. PMID: 12434001

11. Palecek E, Vlk D, Stankova V, Brazda V, Vojtesek B, Hupp TR, et al. Tumor suppressor protein p53 binds preferentially to supercoiled DNA. Oncogene. 1997; 15(18):2201–2209. doi: 10.1038/sj.onc.1201398 PMID: 9393978

12. Kim E, Deppert W. The complex interactions of p53 with target DNA: we learn as we go. Biochem Cell Biol. 2003; 81(3):141–150. doi: 10.1139/o03-046 PMID: 12897847

13. Adamik M, Kejnovská I, Bazantova P, Petr M, Renčiuk D, Vorlíčková M, et al. p53 binds human telomeric G-quadruplex in vitro. Biochimie. 2016.

14. Felsenfeld G, Rich A. Studies on the formation of two- and three-stranded polyribonucleotides. Biochim Biophys Acta. 1957; 26(3):457–468. PMID: 13499402

15. Frank-Kamenetskii MD, Mirkin SM. Triplex DNA structures. Annu Rev Biochem. 1995; 64:65–95. doi: 10.1146/annurev.bi.64.070195.000433 PMID: 7574496

16. Mirkin SM, Frank-Kamenetskii MD. H-DNA and related structures. Annu Rev Biophys Biomol Struct. 1994; 23:541–576. doi: 10.1146/annurev.bb.23.060194.002545 PMID: 7919793

17. Schroth GP, Ho PS. Occurrence of potential cruciform and H-DNA forming sequences in genomic DNA. Nucleic Acids Res. 1995; 23(11):1977–1983. PMID: 7596826

18. Bacolla A, Collins JR, Gold B, Chuzhanova N, Yi M, Stephens RM, et al. Long homopurine*homopyrimidine sequences are characteristic of genes expressed in brain and the pseudoautosomal region. Nucleic Acids Res. 2006; 34(9):2663–2675. doi: 10.1093/nar/gkl354 PMID: 16714445

19. Lexa M, Martinek T, Brazdova M. Uneven Distribution of Potential Triplex Sequences in the Human Genome In Silico Study using the R/Bioconductor Package Triplex. Bioinformatics 2014: Proceedings of the International Conference on Bioinformatics Models, Methods and Algorithms. 2014:80–88.

20. Buske FA, Bauer DC, Mattick JS, Bailey TL. Triplex-Inspector: an analysis tool for triplex-mediated targeting of genomic loci. Bioinformatics (Oxford, England). 2013; 29(15):1895–1897.

21. Vasquez KM, Glazer PM. Triplex-forming oligonucleotides: principles and applications. Q Rev Biophys. 2002; 35(1):89–107. PMID: 11997982

22. Lacroix L, Lacoste J, Reddoch JF, Mergny JL, Levy DD, Seidman MM, et al. Triplex formation by oligonucleotides containing 5-(1-propynyl)-2'-deoxyuridine: decreased magnesium dependence and improved intracellular gene targeting. Biochemistry. 1999; 38(6):1893–1901. doi: 10.1021/bi982290q PMID: 10026270

23. Buske FA, Mattick JS, Bailey TL. Potential in vivo roles of nucleic acid triple-helices. RNA biology. 2011; 8(3):427–439. doi: 10.4161/rna.8.3.14999 PMID: 21525785

24. Guieysse AL, Praseuth D, Helene C. Identification of a triplex DNA-binding protein from human cells. Journal of molecular biology. 1997; 267(2):289–298. doi: 10.1006/jmbi.1997.0884 PMID: 9096226

25. Kiyama R, Camerini-Otero RD. A triplex DNA-binding protein from human cells: purification and characterization. Proceedings of the National Academy of Sciences of the United States of America. 1991; 88 (23):10450–10454. PMID: 1961709

26. Kusic J, Tomic B, Divac A, Kojic S. Human initiation protein Orc4 prefers triple stranded DNA. Molecular biology reports. 2010; 37(5):2317–2322. doi: 10.1007/s11033-009-9735-8 PMID: 19690980

27. Wang G, Carbajal S, Vijg J, DiGiovanni J, Vasquez KM. DNA structure-induced genomic instability in vivo. J Natl Cancer Inst. 2008; 100(24):1815–1817. doi: 10.1093/jnci/djn385 PMID: 19066276

28. Palecek E. Local supercoil-stabilized DNA structures. Crit Rev Biochem Mol Biol. 1991; 26(2):151–226. doi: 10.3109/10409239109081126 PMID: 1914495

29. Brazdova M, Navratilova L, Tichy V, Nemcova K, Lexa M, Hrstka R, et al. Preferential binding of hot spot mutant p53 proteins to supercoiled DNA in vitro and in cells. PLoS One. 2013; 8(3):e59567. doi: 10.1371/journal.pone.0059567 PMID: 23555710

30. Palecek E, Brazda V, Jagelska E, Pecinka P, Karlovska L, Brazdova M. Enhancement of p53 sequence-specific binding by DNA supercoiling. Oncogene. 2004; 23(12):2119–2127. doi: 10.1038/sj.onc.1207324 PMID: 14755248

31. Gohler T, Reimann M, Cherny D, Walter K, Warnecke G, Kim E, et al. Specific interaction of p53 with target binding sites is determined by DNA conformation and is regulated by the C-terminal domain. J Biol Chem. 2002; 277(43):41192–41203. doi: 10.1074/jbc.M202344200 PMID: 12171916

32. Cobb AM, Jackson BR, Kim E, Bond PL, Bowater RP. Sequence-specific and DNA structure-dependent interactions of Escherichia coli MutS and human p53 with DNA. Anal Biochem. 2013; 442(1):51–61. doi: 10.1016/j.ab.2013.07.033 PMID: 23928048

33. Walter K, Warnecke G, Bowater R, Deppert W, Kim E. tumor suppressor p53 binds with high affinity to CTG.CAG trinucleotide repeats and induces topological alterations in mismatched duplexes. The Journal of biological chemistry. 2005; 280(52):42497–42507. doi: 10.1074/jbc.M507038200 PMID: 16230356

34. Palecek E, Brazdova M, Cernocka H, Vlk D, Brazda V, Vojtesek B. Effect of transition metals on binding of p53 protein to supercoiled DNA and to consensus sequence in DNA fragments. Oncogene. 1999; 18(24):3617–3625. doi: 10.1038/sj.onc.1202710 PMID: 10380883

35. Fox KR. Long (dA)n.(dT)n tracts can form intramolecular triplexes under superhelical stress. Nucleic acids research. 1990; 18(18):5387–5391. PMID: 2216711

36. Rohaly G, Chemnitz J, Dehde S, Nunez AM, Heukeshoven J, Deppert W, et al. A novel human p53 isoform is an essential element of the ATR-intra-S phase checkpoint. Cell. 2005; 122(1):21–32. doi: 10.1016/j.cell.2005.04.032 PMID: 16009130

37. Karolchik D, Hinrichs AS, Furey TS, Roskin KM, Sugnet CW, Haussler D, et al. The UCSC Table Browser data retrieval tool. Nucleic acids research. 2004; 32(Database issue):D493–496. doi: 10.1093/nar/gkh103 PMID: 14681465

38. Tebaldi T, Zaccara S, Alessandrini F, Bisio A, Ciribilli Y, Inga A. Whole-genome cartography of p53 response elements ranked on transactivation potential. BMC genomics. 2015; 16:464. doi: 10.1186/s12864-015-1643-9 PMID: 26081755

39. Lexa M, Martinek T, Burgetova I, Kopecek D, Brazdova M. A dynamic programming algorithm for identification of triplex-forming sequences. Bioinformatics (Oxford, England). 2011; 27(18):2510–2517.

40. Szklarczyk D, Franceschini A, Wyder S, Forslund K, Heller D, Huerta-Cepas J, et al. STRING v10: protein-protein interaction networks, integrated over the tree of life. Nucleic acids research. 2015; 43(Database issue):D447–452. doi: 10.1093/nar/gku1003 PMID: 25352553

41. Reimand J, Kull M, Peterson H, Hansen J, Vilo J. g:Profiler—a web-based toolset for functional profiling of gene lists from large-scale experiments. Nucleic acids research. 2007; 35(Web Server issue):W193–200. doi: 10.1093/nar/gkm226 PMID: 17478515

42. Allen MA, Andrysik Z, Dengler VL, Mellert HS, Guarnieri A, Freeman JA, et al. Global analysis of p53-regulated transcription identifies its direct targets and unexpected regulatory mechanisms. Elife. 2014; 3:e02200. doi: 10.7554/eLife.02200 PMID: 24867637

43. Fonseca NA, Marioni J, Brazma A. RNA-Seq gene profiling—a systematic empirical comparison. PLoS One. 2014; 9(9):e107026. doi: 10.1371/journal.pone.0107026 PMID: 25268973

44. Kolesnikov N, Hastings E, Keays M, Melnichuk O, Tang YA, Williams E, et al. ArrayExpress update—simplifying data submissions. Nucleic acids research. 2015; 43(Database issue):D1113–1116. doi: 10.1093/nar/gku1057 PMID: 25361974

45. Kauffmann A, Rayner TF, Parkinson H, Kapushesky M, Lukk M, Brazma A, et al. Importing ArrayExpress datasets into R/Bioconductor. Bioinformatics (Oxford, England). 2009; 25(16):2092–2094.

46. van Noort SJ, van der Werf KO, Eker AP, Wyman C, de Grooth BG, van Hulst NF, et al. Direct visualization of dynamic protein-DNA interactions with a dedicated atomic force microscope. Biophys J. 1998; 74(6):2840–2849. doi: 10.1016/S0006-3495(98)77991-3 PMID: 9635738

47. Pecinka P, Huertas D, Azorin F, Palecek E. Intramolecular TAT triplex in (dA)58.(dT)58. influence of ions. J Biomol Struct Dyn. 1995; 13(1):29–46. doi: 10.1080/07391102.1995.10508819 PMID: 8527029

48. Buzek J, Kuderova A, Pexa T, Stankova V, Lauerova L, Palecek E. Monoclonal antibody against DNA adducts with osmium structural probes. J Biomol Struct Dyn. 1999; 17(1):41–50. doi: 10.1080/07391102.1999.10508339 PMID: 10496420

49. Nejedly K, Chladkova J, Kypr J. Photochemical probing of the B—a conformational transition in a linearized pUC19 DNA and its polylinker region. Biophys Chem. 2007; 125(1):237–246. doi: 10.1016/j.bpc.2006.08.007 PMID: 16962700

50. Sebest P, Brazdova M, Fojta M, Pivonkova H. Differential salt-induced dissociation of the p53 protein complexes with circular and linear plasmid DNA substrates suggest involvement of a sliding mechanism. Int J Mol Sci. 2015; 16(2):3163–3177. doi: 10.3390/ijms16023163 PMID: 25647416

51. Kudoh T, Kimura J, Lu ZG, Miki Y, Yoshida K. D4S234E, a novel p53-responsive gene, induces apoptosis in response to DNA damage. Exp Cell Res. 2010; 316(17):2849–2858. doi: 10.1016/j.yexcr.2010.06.025 PMID: 20599942

52. Sauer M, Bretz AC, Beinoraviciute-Kellner R, Beitzinger M, Burek C, Rosenwald A, et al. C-terminal diversity within the p53 family accounts for differences in DNA binding and transcriptional activity. Nucleic acids research. 2008; 36(6):1900–1912. doi: 10.1093/nar/gkn044 PMID: 18267967

53. Bisio A, De Sanctis V, Del Vescovo V, Denti MA, Jegga AG, Inga A, et al. Identification of new p53 target microRNAs by bioinformatics and functional analysis. BMC Cancer. 2013; 13:552. doi: 10.1186/1471-2407-13-552 PMID: 24256616

54. Janky R, Verfaillie A, Imrichova H, Van de Sande B, Standaert L, Christiaens V, et al. iRegulon: from a gene list to a gene regulatory network using large motif and track collections. PLoS computational biology. 2014; 10(7):e1003731. doi: 10.1371/journal.pcbi.1003731 PMID: 25058159

55. Kracikova M, Akiri G, George A, Sachidanandam R, Aaronson SA. A threshold mechanism mediates p53 cell fate decision between growth arrest and apoptosis. Cell death and differentiation. 2013; 20 (4):576–588. doi: 10.1038/cdd.2012.155 PMID: 23306555

56. Nikulenkov F, Spinnler C, Li H, Tonelli C, Shi Y, Turunen M, et al. Insights into p53 transcriptional function via genome-wide chromatin occupancy and gene expression analysis. Cell death and differentiation. 2012; 19(12):1992–2002. doi: 10.1038/cdd.2012.89 PMID: 22790872

57. Wang B, Niu D, Lam TH, Xiao Z, Ren EC. Mapping the p53 transcriptome universe using p53 natural polymorphs. Cell death and differentiation. 2014; 21(4):521–532. doi: 10.1038/cdd.2013.132 PMID: 24076587

58. Sanchez Y, Segura V, Marin-Bejar O, Athie A, Marchese FP, Gonzalez J, et al. Genome-wide analysis of the human p53 transcriptional network unveils a lncRNA tumour suppressor signature. Nat Commun. 2014; 5:5812. doi: 10.1038/ncomms6812 PMID: 25524025

59. Cer RZ, Bruce KH, Mudunuri US, Yi M, Volfovsky N, Luke BT, et al. Non-B DB: a database of predicted non-B DNA-forming motifs in mammalian genomes. Nucleic Acids Res. 2011; 39(Database issue): D383–391. doi: 10.1093/nar/gkq1170 PMID: 21097885

60. Lopez Castel A, Cleary JD, Pearson CE. Repeat instability as the basis for human diseases and as a potential target for therapy. Nat Rev Mol Cell Biol. 2010; 11(3):165–170. doi: 10.1038/nrm2854 PMID: 20177394

61. Zhao J, Bacolla A, Wang G, Vasquez KM. Non-B DNA structure-induced genetic instability and evolution. Cell Mol Life Sci. 2010; 67(1):43–62. doi: 10.1007/s00018-009-0131-2 PMID: 19727556

62. Goni JR, Vaquerizas JM, Dopazo J, Orozco M. Exploring the reasons for the large density of triplex-forming oligonucleotide target sequences in the human regulatory regions. BMC Genomics. 2006; 7:63. doi: 10.1186/1471-2164-7-63 PMID: 16566817

63. Dudenhoffer C, Rohaly G, Will K, Deppert W, Wiesmuller L. Specific mismatch recognition in heteroduplex intermediates by p53 suggests a role in fidelity control of homologous recombination. Mol Cell Biol. 1998; 18(9):5332–5342. PMID: 9710617

64. Lee S, Elenbaas B, Levine A, Griffith J. p53 and its 14 kDa C-terminal domain recognize primary DNA damage in the form of insertion/deletion mismatches. Cell. 1995; 81(7):1013–1020. PMID: 7600570

65. Quante T, Otto B, Brazdova M, Kejnovska I, Deppert W, Tolstonog GV. Mutant p53 is a transcriptional co-factor that binds to G-rich regulatory regions of active genes and generates transcriptional plasticity. Cell Cycle. 2012; 11(17):3290–3303. doi: 10.4161/cc.21646 PMID: 22894900

66. Subramanian D, Griffith JD. Modulation of p53 binding to Holliday junctions and 3-cytosine bulges by phosphorylation events. Biochemistry. 2005; 44(7):2536–2544. doi: 10.1021/bi048700u PMID: 15709766

67. Brazdova M, Quante T, Togel L, Walter K, Loscher C, Tichy V, et al. Modulation of gene expression in U251 glioblastoma cells by binding of mutant p53 R273H to intronic and intergenic sequences. Nucleic Acids Res. 2009; 37(5):1486–1400. doi: 10.1093/nar/gkn1085 PMID: 19139068

68. Petr M, Helma R, Polaskova A, Krejci A, Dvorakova Z, Kejnovska I, et al. Wild-type p53 binds to MYC promoter G-quadruplex. Biosci Rep. 2016; 36(5).

69. Kim H, Kim K, Choi J, Heo K, Baek HJ, Roeder RG, et al. p53 requires an intact C-terminal domain for DNA binding and transactivation. Journal of molecular biology. 2012; 415(5):843–854. doi: 10.1016/j.jmb.2011.12.001 PMID: 22178617

70. Laptenko O, Shiff I, Freed-Pastor W, Zupnick A, Mattia M, Freulich E, et al. The p53 C terminus controls site-specific DNA binding and promotes structural changes within the central DNA binding domain. Mol Cell. 2015; 57(6):1034–1046. doi: 10.1016/j.molcel.2015.02.015 PMID: 25794615

71. Laptenko O, Tong DR, Manfredi J, Prives C. The Tail That Wags the Dog: How the Disordered C-Terminal Domain Controls the Transcriptional Activities of the p53 Tumor-Suppressor Protein. Trends Biochem Sci. 2016.

72. Friedler A, Veprintsev DB, Freund SM, von Glos KI, Fersht AR. Modulation of binding of DNA to the C-terminal domain of p53 by acetylation. Structure. 2005; 13(4):629–636. doi: 10.1016/j.str.2005.01.020 PMID: 15837201

73. McKinney K, Prives C. Efficient specific DNA binding by p53 requires both its central and C-terminal domains as revealed by studies with high-mobility group 1 protein. Mol Cell Biol. 2002; 22(19):6797–6808. doi: 10.1128/MCB.22.19.6797-6808.2002 PMID: 12215537

74. McKinney K, Mattia M, Gottifredi V, Prives C. p53 linear diffusion along DNA requires its C terminus. Mol Cell. 2004; 16(3):413–424. doi: 10.1016/j.molcel.2004.09.032 PMID: 15525514

75. Kenzelmann Broz D, Spano Mello S, Bieging KT, Jiang D, Dusek RL, Brady CA, et al. Global genomic profiling reveals an extensive p53-regulated autophagy program contributing to key p53 responses. Genes Dev. 2013; 27(9):1016–1031. doi: 10.1101/gad.212282.112 PMID: 23651856

76. Tutton S, Azzam GA, Stong N, Vladimirova O, Wiedmer A, Monteith JA, et al. Subtelomeric p53 binding prevents accumulation of DNA damage at human telomeres. Embo J. 2016; 35(2):193–207. doi: 10.15252/embj.201490880 PMID: 26658110

77. Kaushik Tiwari M, Adaku N, Peart N, Rogers FA. Triplex structures induce DNA double strand breaks via replication fork collapse in NER deficient cells. Nucleic acids research. 2016; 44(16):7742–7754. doi: 10.1093/nar/gkw515 PMID: 27298253

78. Hampp S, Kiessling T, Buechle K, Mansilla SF, Thomale J, Rall M, et al. DNA damage tolerance pathway involving DNA polymerase iota and the tumor suppressor p53 regulates DNA replication fork progression. Proceedings of the National Academy of Sciences of the United States of America. 2016; 113 (30):E4311–4319. doi: 10.1073/pnas.1605828113 PMID: 27407148

79. Reed M, Woelker B, Wang P, Wang Y, Anderson ME, Tegtmeyer P. The C-terminal domain of p53 recognizes DNA damaged by ionizing radiation. Proceedings of the National Academy of Sciences of the United States of America. 1995; 92(21):9455–9459. PMID: 7568153

80. Bacolla A, Wang G, Vasquez KM. New Perspectives on DNA and RNA Triplexes As Effectors of Biological Activity. PLoS Genet. 2015; 11(12):e1005696. doi: 10.1371/journal.pgen.1005696 PMID: 26700634

81. Wang G, Vasquez KM. Models for chromosomal replication-independent non-B DNA structure-induced genetic instability. Mol Carcinog. 2009; 48(4):286–298. doi: 10.1002/mc.20508 PMID: 19123200

82. Collavin L, Lunardi A, Del Sal G. p53-family proteins and their regulators: hubs and spokes in tumor suppression. Cell Death Differ. 2010; 17(6):901–911. doi: 10.1038/cdd.2010.35 PMID: 20379196

83. Menendez D, Nguyen TA, Freudenberg JM, Mathew VJ, Anderson CW, Jothi R, et al. Diverse stresses dramatically alter genome-wide p53 binding and transactivation landscape in human cancer cells. Nucleic acids research. 2013; 41(15):7286–7301. doi: 10.1093/nar/gkt504 PMID: 23775793

84. Liu X, Tan Y, Zhang C, Zhang Y, Zhang L, Ren P, et al. NAT10 regulates p53 activation through acetylating p53 at K120 and ubiquitinating Mdm2. EMBO Rep. 2016; 17(3):349–366. doi: 10.15252/embr.201540505 PMID: 26882543

85. Younger ST, Kenzelmann-Broz D, Jung H, Attardi LD, Rinn JL. Integrative genomic analysis reveals widespread enhancer regulation by p53 in response to DNA damage. Nucleic acids research. 2015; 43 (9):4447–4462. doi: 10.1093/nar/gkv284 PMID: 25883152

86. Melo CA, Drost J, Wijchers PJ, van de Werken H, de Wit E, Oude Vrielink JA, et al. eRNAs are required for p53-dependent enhancer activity and gene transcription. Mol Cell. 2013; 49(3):524–535. doi: 10.1016/j.molcel.2012.11.021 PMID: 23273978

87. Cherny DI, Brazdova M, Palecek J, Palecek E, Jovin TM. Sequestering of p53 into DNA-protein filaments revealed by electron microscopy. Biophys Chem. 2005; 114(2–3):261–271. doi: 10.1016/j.bpc.2004.12.042 PMID: 15829361

A.6 Paper VI

SoluProt: prediction of soluble protein expression in Escherichia coli

Bioinformatics, 37(1), 2021, 23–28
doi: 10.1093/bioinformatics/btaa1102
Advance Access Publication Date: 8 January 2021
Original Paper

Sequence analysis

SoluProt: prediction of soluble protein expression in *Escherichia coli*

Jiri Hon[1,2,3], Martin Marusiak[3], Tomas Martinek[3], Antonin Kunka[1,2], Jaroslav Zendulka[3], David Bednar[1,2,*] and Jiri Damborsky [1,2,*]

[1]Loschmidt Laboratories, Centre for Toxic Compounds in the Environment RECETOX and Department of Experimental Biology, Faculty of Science, Masaryk University, Brno 625 00, Czech Republic, [2]International Clinical Research Center, St. Anne's University Hospital Brno, Brno 656 91, Czech Republic and [3]IT4Innovations Centre of Excellence, Faculty of Information Technology, Brno University of Technology, Brno 612 66, Czech Republic

*To whom correspondence should be addressed.

Associate Editor: Jinbo Xu

Received on October 3, 2020; revised on December 5, 2020; editorial decision on December 24, 2020; accepted on December 28, 2020

Abstract

Motivation: Poor protein solubility hinders the production of many therapeutic and industrially useful proteins. Experimental efforts to increase solubility are plagued by low success rates and often reduce biological activity. Computational prediction of protein expressibility and solubility in *Escherichia coli* using only sequence information could reduce the cost of experimental studies by enabling prioritization of highly soluble proteins.

Results: A new tool for sequence-based prediction of soluble protein expression in *E.coli*, SoluProt, was created using the gradient boosting machine technique with the TargetTrack database as a training set. When evaluated against a balanced independent test set derived from the NESG database, SoluProt's accuracy of 58.5% and AUC of 0.62 exceeded those of a suite of alternative solubility prediction tools. There is also evidence that it could significantly increase the success rate of experimental protein studies. SoluProt is freely available as a standalone program and a user-friendly webserver at https://loschmidt.chemi.muni.cz/soluprot/.

Availability and implementation: https://loschmidt.chemi.muni.cz/soluprot/.

Contact: jiri@chemi.muni.cz

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Low protein solubility causes severe problems in protein science and industry; insufficient protein solubility is probably the most common cause of failure of protein production pipelines. The importance of solubility is underlined by the findings of the large-scale Protein Structure Initiative (PSI) project (Berman *et al.*, 2017), which sought to produce thousands of protein sequences from different organisms, crystallize them and resolve their tertiary structure. Unfortunately, in most cases it proved impossible to produce the target proteins in soluble form. The inherent low solubility of natural enzymes also limits the success of emerging high-throughput pipelines that explore protein databases to identify novel enzymes with diverse functions (Hon *et al.*, 2020; Vanacek *et al.*, 2018). Given the rapid growth of protein sequence databases driven by the capabilities of next-generation sequencing technologies, there is an urgent need to focus only on potentially soluble targets to avoid wasting resources on hard-to-produce orthologs. Solubility is thus a key attribute when choosing protein targets for experimental characterization (Vanacek *et al.*, 2018). Strictly speaking, solubility is a thermodynamic parameter defined as the protein's concentration in a saturated solution in equilibrium with a solid phase under specific conditions. However, it is challenging to quantitatively measure the solubility of large sets of proteins (Kramer *et al.*, 2012), so there is little quantitative experimental data on protein solubility. Moreover, this definition of solubility is too narrow to encompass many of the practical problems that may occur during protein production with common expression systems. Therefore, inspired by existing tools (Supplementary Table S1) (Agostini *et al.*, 2014; Khurana *et al.*, 2018; Raimondi *et al.*, 2020; Smialowski *et al.*, 2012), available data (Berman *et al.*, 2017) and laboratory practice, we use a slightly extended definition of protein solubility in this work. Specifically, by solubility, we mean the probability of soluble protein (over)expression in *Escherichia coli* cells. The difference from the classical thermodynamic solubility is in the perception of the insoluble class. We assume that insoluble proteins were either not expressed or were expressed in the insoluble form.

Solubility depends on many extrinsic and intrinsic factors. Extrinsic factors are dictated by the choice of expression system and the experimental conditions used in protein production. Expression systems may be either *in vivo* or *in vitro* (Carlson *et al.*, 2012; Rosano and Ceccarelli, 2014). *In vivo* protein expression is induced inside living cells of a host organism, whereas *in vitro* expression relies on the use of cell-free translational systems. Solubility can be increased by adjusting extrinsic solubility factors, especially by using different mutated host strains, codon optimization, coexpression of chaperones and foldases, lowering cultivation temperatures and adding suitable fusion partners (Costa *et al.*, 2014). However, tuning the expression system or experimental conditions is not always sufficient to confer solubility, and is not feasible in high-throughput protein production pipelines. If extrinsic factors cannot be varied, protein solubility will depend only on the intrinsic properties of the protein sequence. Unfortunately, the relationship between a protein's sequence and its solubility is poorly understood, mainly due to a lack of reproducible quantitative solubility measurements (Kramer *et al.*, 2012). Recent protein engineering studies suggest that charged amino acids on the protein surface are key intrinsic determinants of solubility (Carballo-Amador *et al.*, 2019; Chan *et al.*, 2013; Sankar *et al.*, 2018). However, this knowledge cannot be directly used for solubility prediction due to a lack of structural data. Despite the continuous growth of structural databases (Burley *et al.*, 2019), the structures of proteins of interest are generally unknown, and the limited availability of template structures prevents their accurate computational prediction.

The simultaneous effects of extrinsic and intrinsic factors make solubility prediction challenging. For example, the prediction of solubility from sequence data using machine learning is hindered by the high level of noise in typical training datasets due to the influence of diverse extrinsic variables. Because the molecular mechanisms governing protein solubility are poorly understood, recent solubility prediction tools rely heavily on statistical analysis and machine learning, using previously reported experimental data to train and validate model parameters. One of the most widely used data sources is the TargetTrack database (Berman *et al.*, 2017), formerly known as PepcDB or TargetDB, which integrates information from the Protein Structure Initiative projects. This database contains data from over 900 000 protein crystallization trials involving almost 300 000 unique protein sequences, which are referred to as targets. The database does not contain solubility data per se, but target proteins can be considered soluble if they were successfully purified in the experimental trials. A major limitation of this database is the low quality of its annotations. For example, reasons for failure are generally not provided for unsuccessful crystallization attempts. Therefore, it is impossible to distinguish failures due to insolubility from failures due to other problems later in the experimental pipeline. Second, the experimental protocols used for protein production and crystallization are described in free text with no internal structure, making it hard to automatically extract information about experimental conditions and expression systems for a given target. Filtering is therefore needed to reduce noise before using the TargetTrack data for model training. However, the application of stringent filtering rules to the target annotations can dramatically reduce the number of usable records.

eSOL is another well-known and commonly used solubility database (Niwa *et al.*, 2009, 2012) that contains experimentally measured solubilities for over 3 000 *E.coli* proteins produced in the PURE (Shimizu *et al.*, 2001) cell-free expression system. eSOL is an impressive collection of highly homogenous data but has its own limitations. First, it only contains data on proteins originating from *E.coli*. Second, it has relatively little negative data; adding the three main cytosolic *E.coli* chaperones (TF, DnaKJE and GroEL/GroES) to the PURE expression system reduced the number of insoluble proteins from 788 to 24 (Niwa *et al.*, 2012). eSOL is a valuable source of exact solubility data that were generated using a robust pipeline and provide a good quantitative measure of thermodynamic solubility. However, these data cannot be used to assess solubility according to our expanded definition, which also encompasses expressibility.

The relationship between protein sequence and solubility has been studied for over 30 years, leading to the development of several predictive models and software tools. There are 11 such models or tools that use definitions of solubility like that described above and take protein sequences as their sole input. These are the revised Wilkinson-Harrison model (rWH) (Davis *et al.*, 1999; Wilkinson and Harrison, 1991), SOLpro (Magnan *et al.*, 2009), RPSP (Diaz *et al.*, 2010), PROSO II (Smialowski *et al.*, 2012), ccSOL omics (Agostini *et al.*, 2012, 2014), ESPRESSO (Hirose and Noguchi, 2013), CamSol (Sormanni *et al.*, 2015), Protein-Sol (Hebditch *et al.*, 2017), DeepSol (Khurana *et al.*, 2018), SKADE (Raimondi *et al.*, 2020) and the Solubility-weighted index (SWI) (Bhandari *et al.*, 2020). However, the accuracy of these tools is limited, and there is clear room for improvement. Additionally, these tools exhibit poor generality when used to make predictions based on previously unseen data. A comprehensive review of advances in solubility prediction, including predictors that use protein structures as inputs, was published recently (Musil *et al.*, 2019). Here, we present a novel machine learning based tool, SoluProt, for predicting soluble expression from protein sequence data. SoluProt benefits from thorough dataset pre-processing and predicts soluble expression more accurately than previously reported methods.

2 SoluProt training and test set

We used the TargetTrack database to build the *SoluProt training set*. Since this database does not directly provide solubility information, we inferred solubility computationally, using an approach similar to those adopted previously (Magnan *et al.*, 2009; Smialowski *et al.*, 2012). A protein was considered *soluble* if it was recorded as having reached a soluble experimental state or any subsequent state requiring soluble expression (Supplementary Table S2). If failed expression or purification was mentioned in the experiment record's stop status, the protein was labeled *insoluble*. In contrast to a previous approach (Smialowski *et al.*, 2012), we required an explicit stop status relating to insolubility to reduce the frequency of incorrect classification of insoluble sequences. To improve the quality of the training set, we also performed several additional steps to clean the data.

Most importantly, we performed keyword matching combined with manual checking of TargetTrack annotations to extract only proteins expressed in the most common host organism, *E.coli*. This was necessary because a protein soluble in one organism might be insoluble in another. By focusing solely on the most common expression system, we reduced the noise in the training data. We also used specific keywords to search the unstructured descriptions of experimental protocols provided in the TargetTrack database (Supplementary Table S3). Generic search phrases like '*E.coli*' or '*Escherichia coli*' were used to identify potential *E.coli* related protocols. These protocols were then manually checked and confirmed (Supplementary Table S4). A full list of 248 TargetTrack protocols signifying expression in *E.coli* is available at the SoluProt website.

We next identified transmembrane proteins in the dataset based on direct annotations from the TargetTrack database and predictions generated using TOPCONS (Tsirigos *et al.*, 2015) with default settings. The transmembrane proteins were then removed, along with sequences shorter than 20 amino acids, and sequences with undefined residues. We also removed sequences that had been classified as insoluble but for which a protein structure was available in the Protein Data Bank (PDB) (Berman, 2000). To this end, we compiled an *E.coli* PDB subset containing sequences of proteins whose structures had been solved by NMR or X-ray crystallography and which had been expressed in *E.coli* according to the PDB annotations (64 416 sequences, downloaded April 4, 2018). Because both NMR and X-ray crystallography require soluble proteins, any protein in this PDB subset can be considered soluble in *E.coli*. This step reflects advances in molecular biology: methodological developments have made it possible to produce and crystallize some proteins that were previously considered insoluble.

Finally, we reduced the sequence redundancy in the training set by clustering to 25% identity using MMseqs2 (Steinegger and

Söding, 2017) and retaining only representative sequences from each cluster. This was done separately for positive and negative samples to avoid simplifying the prediction problem. We balanced the number of soluble and insoluble samples such that both classes were equally represented. Additionally, we balanced the sequence length distribution so that length alone would not play a dominant role in the predictions. Sequence length correlates with protein solubility—larger proteins are usually less soluble. However, we wanted to suppress its influence in the model because we anticipate that SoluProt would mainly be used to prioritize proteins of similar lengths, usually from a single protein family. A typical expected use case is that of the EnzymeMiner web server (Hon *et al.*, 2020) for automated mining of soluble enzymes. A prediction model relying heavily on sequence length would not perform well in this use case.

The *SoluProt test set* was built from a dataset generated by the North East Structural Consortium (NESG), which represents 9644 proteins expressed in *E.coli* using a unified production pipeline (Price *et al.*, 2011). The dataset contains two integer scores ranging from 0 to 5 for each target, indicating the protein's level of expression and the soluble fraction recovery. The reproducibility of the experimental results in the dataset was validated by performing repeat measurements for selected targets. The NESG dataset targets are included in the TargetTrack database because the NESG participated in the PSI project. However, the expression and solubility levels from the NESG dataset were not included in the TargetTrack database; instead, they were provided to us directly by the authors of the original study (W. Nicholson Price II, personal communication). The high consistency and quality of the NESG dataset make it suitable for benchmarking purposes. We processed the NESG dataset using the same procedure as the training set, although the computational solubility derivation and expression system filtration steps were omitted because they were pointless in this case. Instead, we transformed the solubility levels into binary classes: all proteins with a solubility level of 1 or above were considered soluble and all others insoluble.

Finally, we ensured that no pair consisting of a sequence from the test set and a sequence from the training set had a global sequence identity above 25% as calculated using the USEARCH software (Edgar, 2010). This made the test set more independent because it ensured that predictions were not validated against data similar to those used during training. In total, 11 436 protein sequences remained in the *SoluProt training set* and 3 100 in the independent *SoluProt test set*. Both datasets had equal numbers of soluble and insoluble samples with balanced sequence length distributions (Supplementary Fig. S1). The datasets are available at the SoluProt website. The dataset construction steps are summarized in Supplementary Table S5.

3 Prediction model

The SoluProt predictor is implemented in Python using scikit-learn (Pedregosa *et al.*, 2011), Biopython (Cock *et al.*, 2009) and pandas (McKinney, 2010) libraries. We used a gradient boosting machine (GBM) (Friedman, 2001) to generate the predictive model. Prediction features were selected from a set of 251 sequence characteristics that were divided into eight groups: (i) single amino acid content (20 features), (ii) amino acid dimer content (210 features), (iii), sequence physicochemical features (12 features, Supplementary Table S6), (iv) average flexibility as computed by DynaMine (Cilia *et al.*, 2014) (1 feature), (v) secondary structure content as predicted by FESS (Piovesan *et al.*, 2017) (3 features), (vi) average disorder as predicted by ESPRITZ (Walsh *et al.*, 2012) (1 feature), (vii) content of amino acids in transmembrane helices as predicted by TMHMM (Krogh *et al.*, 2001) (3 features) and (viii) maximum identity to the *E.coli* PDB subset as calculated using USEARCH (1 feature). All sequences equal to any sequence from the test set were excluded from the *E.coli* PDB subset for the calculation of maximum identity. The objective was to eliminate even the indirect presence of test set sequences from model training. We standardized all features by subtracting the mean and scaling to unit variance. The means and variances were calculated using the training set.

We removed correlated features in two steps. First, we fitted a GBM with default parameters using the full training set and all features. Second, we calculated Pearson's correlation coefficient for each pair of features. If the correlation between any two features exceeded 0.75, we removed the feature with the lesser importance in the fitted GBM model. We also removed irrelevant features using LASSO (Tibshirani, 1996). LASSO's alpha parameter was optimized to maximize the mean AUC of the GBM model with default parameters over 5-fold cross-validation. The alpha parameter was varied between 0.08 to 0 with a step size of 6.25×10^{-4}; its optimal value was 0.005. In total, 96 features were selected for inclusion in the predictive model (Supplementary Table S7). The DynaMine, FESS and ESPRITZ features were not included in the final feature set.

We next optimized the hyperparameters of the GBM model, using an iterative 7-stage strategy to maximize the mean AUC over 5-fold cross-validation using the training set (Supplementary Table S8). In each stage, one or two parameters were optimized using grid search; other parameters were left either at their final values from the previous stages or at the default value if the parameter had not yet been optimized. The best GBM model achieved mean AUC values of 0.85 ± 0.003 for the training part and 0.72 ± 0.02 for the validation part. Overall, the feature selection and hyperparameter optimization had little effect on the mean AUC: without these measures, the mean AUC values for the training and validation sets were 0.83 ± 0.003 and 0.72 ± 0.02, respectively. The main benefit of the feature selection and parameter tuning steps was that they reduced the number of features and thus made the feature calculation step roughly two times faster.

Finally, we used the best GBM hyperparameters to train the final SoluProt model using the full training set. The resulting model had an AUC of 0.84 and an accuracy of 76% for the full training set. The five most important features according to the GBM are: (i) maximum identity to the *E.coli* PDB subset (14.5%), (ii) isoelectric point (6.2%), (iii) predicted number of amino acids in transmembrane helices in the first sixty amino acids of the protein (4.2%), (iv) lysine content (4.0%) and (v) glutamine content (3.5%) (Supplementary Table S7).

4 Performance evaluation and comparison

We used the SoluProt test set to evaluate and compare SoluProt to 11 previously published tools. The evaluation relied on both threshold-independent (area under the ROC curve) and threshold-dependent metrics (accuracy, Matthew's correlation coefficient and confusion matrices). For the threshold-dependent metrics, we applied a threshold of 0.5 or the thresholds recommended by the authors of the corresponding method (Table 1). SoluProt achieved the highest accuracy (58.5%) and the greatest AUC (0.62) of the

Table 1. Performance of various solubility predictors using the balanced SoluProt test set of 3100 sequences

Method	AUC	T	ACC	MCC	TP	TN	FP	FN
SoluProt	0.62	0.50	58.5%	0.17	939	873	677	611
PROSO II	0.60	0.60	58.0%	0.17	630	1167	383	920
SWI	0.60	0.50	55.9%	0.13	1206	527	1023	344
CamSol	0.57	1.00	54.1%	0.08	676	1001	549	874
ESPRESSO	0.56	0.50	53.8%	0.08	1003	664	886	547
rWH	0.55	0.50	54.0%	0.08	670	1005	545	880
DeepSol	0.55	0.50	52.9%	0.09	230	1409	141	1320
Protein-Sol	0.54	0.45	51.6%	0.03	1056	544	1006	494
SOLpro	0.53	0.50	52.0%	0.04	654	959	591	896
SKADE	0.51	0.50	49.2%	–0.03	159	1366	184	1391
ccSOL omics	0.51	0.50	50.8%	0.02	884	690	860	666
RPSP	0.50	0.50	49.8%	0.00	501	1044	506	1049

Note: The different definitions of solubility and target expression system (Supplementary Table S1) should be considered when comparing the performance of individual tools.

AUC—area under the ROC curve, T—threshold for the soluble class, ACC—accuracy, MCC—Matthew's correlation coefficient, TP—true positives, TN—true negatives, FP—false positives, FN—false negatives.

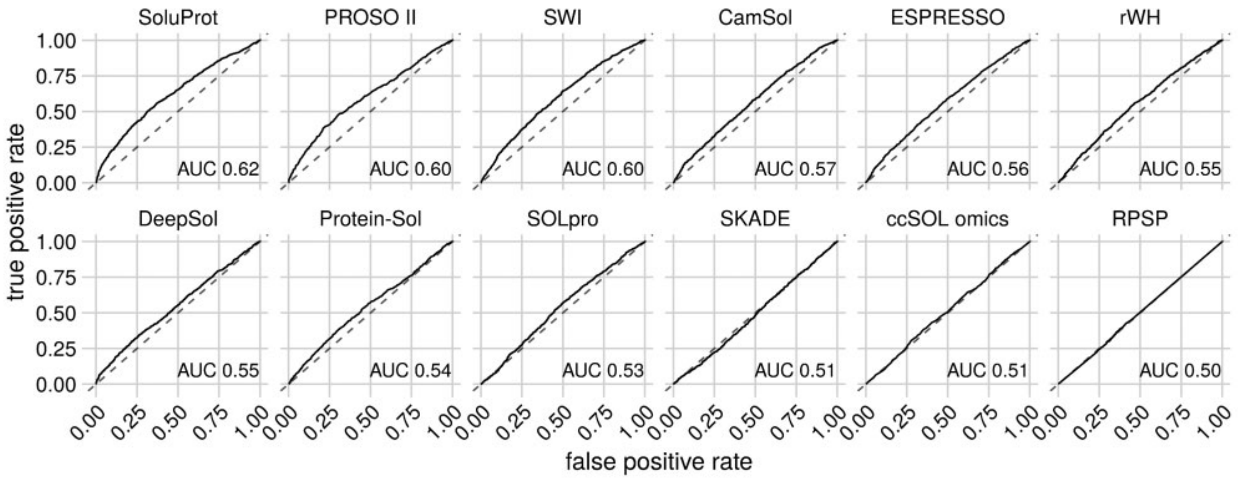


Fig. 1. Receiver operating curves (ROC) calculated for the balanced SoluProt test set of 3100 sequences. The predictors are ordered by the area under the receiver operating curve (AUC)

Table 2. Overlaps between the SoluProt test set and available training sets

Dataset	Size	Test set overlap	TP	TN	FP	FN
PROSO II initial	129643	2952 (95.2%)	951	1437	50	514
DeepSol/SKADE	69420	2294 (74.0%)	737	1130	67	360
SWI	12216	820 (26.5%)	537	210	53	20
SOLpro	17408	480 (15.5%)	178	120	39	143

Note: Two sequences were considered identical if their global sequence identity reported by USEARCH was 100%. Differences in solubility annotations for identical sequences were quantified using confusion matrix terms (TP, TN, FP and FN). The solubility annotations of the SoluProt test set are assumed to reflect the true solubilities of the proteins.

TP—true positives, TN—true negatives, FP—false positives, FN—false negatives. [a] DeepSol and SKADE share the same training set.

tested tools when evaluated against the SoluProt test set (Table 1 and Fig. 1),followed by PROSO II and SWI.

While the SoluProt test set is independent of the SoluProt training set, other tools' training sets might overlap with our test set. Therefore, we compared the SoluProt test set to the training sets of DeepSol, SKADE, SWI and SOLpro to quantify their overlaps (Table 2). DeepSol and SKADE have a common training set, which showed the largest overlap (74.0%), followed by the SWI training set (26.5%) and the SOLpro training set (15.5%). SWI benefits from the overlap; it was the third-best tool in our comparison. DeepSol and SKADE ranked 7th and 12th by accuracy with respect to the SoluProt test set despite having the greatest proportion of test sequences in their training set. This comparatively poor performance can be partly explained by differences in solubility annotations between the DeepSol training set and the SoluProt test set (Table 2): 360 (11.6% of the total) sequences annotated as insoluble in the DeepSol training set were annotated as soluble in the SoluProt test set. The total number of disagreements (the sum of false positives and false negatives) ranged from 336 to 551, depending on the binarization threshold applied to the SoluProt test set (Supplementary Table S9). No training set was published for PROSO II; only an initial set of soluble and insoluble sequences without pre-processing is available. However, the initial set exhibits 95.2% overlap with the SoluProt test set. Therefore, we expect the overlap of the PROSO II training set to also be very high, like the DeepSol training set. Unfortunately, the training sets of other previously developed tools have not been published, preventing a more comprehensive comparison.

The absolute accuracy of the available solubility prediction tools is low (below 60%), so there is clearly room for improvement. Nevertheless, SoluProt and other tools can be useful for protein sequence prioritization (Fig. 2), i.e. for selecting a small number of sequences for in-depth experimental characterization from a large database of several hundreds or thousands of sequences. Specifically, predicted solubility values can be used to select a limited number of high-scoring protein sequences. For example, if we use SoluProt predictions to order the SoluProt test set and remove all sequences bar the 10% with the highest scores, we get 232 true positives, i.e. 49.7% more true positives than would be expected with blind selection (155 true positives). This shows that despite their limited accuracy, current solubility predictors are valuable for protein sequence prioritization and can increase the success rate of experimental protein studies.

5 Conclusions

We have developed a novel method and software tool, SoluProt, for sequence-based prediction of soluble protein expression in *E.coli*. The tool simultaneously predicts the solubility and expressibility of the proteins under consideration. SoluProt achieved a higher accuracy (58.5%) and AUC (0.62) than a suite of alternative solubility prediction tools when evaluated using the balanced independent SoluProt test set of 3100 sequences. PROSO II, SWI and CamSol were the next best tools, achieving accuracies of 58.0%, 55.9% and 54.1%, respectively. SoluProt also performed well in protein prioritization. The main strengths of SoluProt are that it was trained using a dataset generated by thorough pre-processing of the noisy TargetTrack data, and was validated using a high-quality independent test set.

Surprisingly, the recently reported DeepSol (Khurana *et al.*, 2018) and SKADE (Raimondi *et al.*, 2020) tools, which are based on deep learning methods, performed worse than the simpler and mostly older methods PROSO II (Smialowski *et al.*, 2012), SWI (Bhandari *et al.*, 2020) and CamSol (Sormanni *et al.*, 2015) in our comparison. This may be partly due to the overlap of their training set with our test set and disagreements between these sets with respect to the solubility of certain sequences.

The SoluProt predictor is available via a user-friendly web server or as a standalone software package at https://loschmidt.chemi.muni.cz/soluprot/. The SoluProt web server has already predicted the solubility of over 4700 unique protein sequences in ten months since its launch in February 2020. It has also been integrated into the web server EnzymeMiner (Hon *et al.*, 2020) for automated

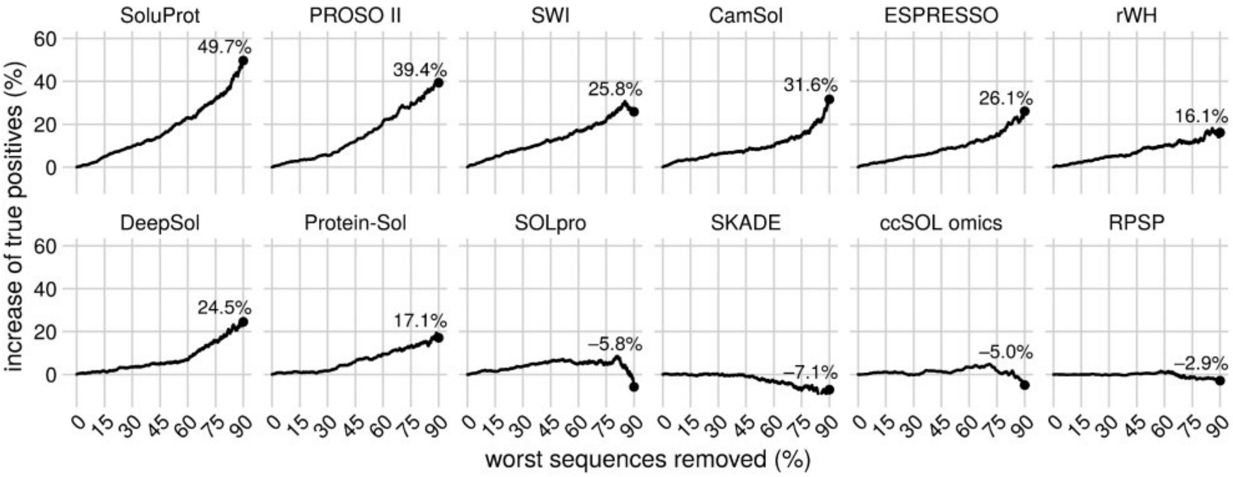


Fig. 2. Increases in the number of true positives resulting from sequence prioritization using the tested solubility prediction tools. The SoluProt test set sequences were ordered by predicted solubility based on each predictor's output, and a variable percentage of the sequences with the worst predicted solubility was then removed. The increase in the number of true positives was then calculated relative to a baseline random selection. For example, upon randomly removing 90% of the test set sequences (2790 samples), we would expect half of the remaining 310 sequences to be true positives

mining of novel soluble enzymes from protein databases (https://loschmidt.chemi.muni.cz/enzymeminer/).

Funding

This work was supported by Czech Ministry of Education [857560, 02.1.01/0.0/0.0/18_046/0015975, CZ.02.1.01/0.0/0.0/16_026/0008451, LQ1602]; Czech Grant Agency (20-15915Y); European Commission [857560, 720776, 814418]; and AI Methods for Cybersecurity and Control Systems project of the Brno University of Technology [FIT-S-20-6293]. Computational resources were supplied by the project 'e-Infrastruktura CZ' [e-INFRA LM2018140] and by the ELIXIR-CZ [LM2018131]. Funding for open access charge: Czech Ministry of Education.

Conflict of Interest: none declared.

References

Agostini,F. *et al.* (2014) ccSOL omics: a webserver for solubility prediction of endogenous and heterologous expression in *Escherichia coli*. *Bioinformatics*, **30**, 2975–2977.

Agostini,F. *et al.* (2012) Sequence-based prediction of protein solubility. *J. Mol. Biol.*, **421**, 237–241.

Berman,H.M. *et al.* (2017) Protein Structure Initiative – TargetTrack 2000-2017 – all data files. *Zenodo*. doi:10.5281/zenodo.821654.

Berman,H.M. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.

Bhandari,B.K. *et al.* (2020) Solubility-Weighted Index: fast and accurate prediction of protein solubility. *Bioinformatics*, **36**, 4691–4698.

Burley,S.K. *et al.* (2019) RCSB Protein Data Bank: biological macromolecular structures enabling research and education in fundamental biology, biomedicine, biotechnology and energy. *Nucleic Acids Res.*, **47**, D464.

Carballo-Amador,M.A. *et al.* (2019) Surface patches on recombinant erythropoietin predict protein solubility: engineering proteins to minimise aggregation. *BMC Biotechnology*, **19**, 26.

Carlson,E.D. *et al.* (2012) Cell-free protein synthesis: applications come of age. *Biotechnol. Adv.*, **30**, 1185–1194.

Chan,P. *et al.* (2013) Soluble expression of proteins correlates with a lack of positively-charged surface. *Sci. Rep.*, **3**, 3333.

Cilia,E. *et al.* (2014) The DynaMine webserver: predicting protein dynamics from sequence. *Nucleic Acids Res.*, **42**, W264–W270.

Cock,P.J.A. *et al.* (2009) Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, **25**, 1422–1423.

Costa,S. *et al.* (2014) Fusion tags for protein solubility, purification and immunogenicity in *Escherichia coli*: the novel Fh8 system. *Front. Microbiol.*, **5**, 63.

Davis,G.D. *et al.* (1999) New fusion protein systems designed to give soluble expression in Escherichia coli. *Biotechnol. Bioeng.*, **65**, 382–388.

Diaz,A.A. *et al.* (2010) Prediction of protein solubility in *Escherichia coli* using logistic regression. *Biotechnol. Bioeng.*, **105**, 374–383.

Edgar,R.C. (2010) Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*, **26**, 2460–2461.

Friedman,J.H. (2001) Greedy function approximation: a gradient boosting machine. *Ann. Stat.*, **29**, 1189–1232.

Hebditch,M. *et al.* (2017) Protein–Sol: a web tool for predicting protein solubility from sequence. *Bioinformatics*, **33**, 3098–3100.

Hirose,S. and Noguchi,T. (2013) ESPRESSO: a system for estimating protein expression and solubility in protein expression systems. *Proteomics*, **13**, 1444–1456.

Hon,J. *et al.* (2020) EnzymeMiner: automated mining of soluble enzymes with diverse structures, catalytic properties and stabilities. *Nucleic Acids Res.*, **48**, W104–W109.

Khurana,S. *et al.* (2018) DeepSol: a deep learning framework for sequence-based protein solubility prediction. *Bioinformatics*, **34**, 2605–2613.

Kramer,R.M. *et al.* (2012) Toward a molecular understanding of protein solubility: increased negative surface charge correlates with increased solubility. *Biophys. J.*, **102**, 1907–1915.

Krogh,A. *et al.* (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol.*, **305**, 567–580.

Magnan,C.N. *et al.* (2009) SOLpro: accurate sequence-based prediction of protein solubility. *Bioinformatics*, **25**, 2200–2207.

McKinney,W. (2010) Data Structures for Statistical Computing in Python. In: *Proceedings of the 9th Python in Science Conference*. SciPy Organizers, Austin, Texas, pp. 56–61.

Musil,M. *et al.* (2019) Computational design of stable and soluble biocatalysts. *ACS Catal.*, **9**, 1033–1054.

Niwa,T. *et al.* (2009) Bimodal protein solubility distribution revealed by an aggregation analysis of the entire ensemble of *Escherichia coli* proteins. *Proc. Natl. Acad. Sci. USA*, **106**, 4201–4206.

Niwa,T. *et al.* (2012) Global analysis of chaperone effects using a reconstituted cell-free translation system. *Proc. Natl. Acad. Sci. USA*, **109**, 8937–8942.

Pedregosa,F. *et al.* (2011) Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.*, **12**, 2825–2830.

Piovesan,D. *et al.* (2017) FELLS: fast estimator of latent local structure. *Bioinformatics*, **33**, 1889–1891.

Price,W.N. *et al.* (2011) Large-scale experimental studies show unexpected amino acid effects on protein expression and solubility in vivo in *E. coli*. *Microb. Inf. Exp.*, **1**, 6.

Raimondi,D. *et al.* (2020) Insight into the protein solubility driving forces with neural attention. *PLoS Comput. Biol.*, **16**, e1007722.

Rosano,G.L. and Ceccarelli,E.A. (2014) Recombinant protein expression in *Escherichia coli*: advances and challenges. *Front. Microbiol.*, **5**, 172.

Sankar,K. *et al.* (2018) AggScore: prediction of aggregation-prone regions in proteins based on the distribution of surface patches. *Proteins*, **86**, 1147–1156.

Shimizu,Y. *et al.* (2001) Cell-free translation reconstituted with purified components. *Nat. Biotechnol.*, **19**, 751–755.

Smialowski,P. *et al.* (2012) PROSO II - a new method for protein solubility prediction. *FEBS J.*, **279**, 2192–2200.

Sormanni,P. *et al.* (2015) The CamSol method of rational design of protein mutants with enhanced solubility. *J. Mol. Biol.*, **427**, 478–490.

Steinegger,M. and Söding,J. (2017) MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat. Biotechnol.*, **35**, 1026–1028.

Tibshirani,R. (1996) Regression shrinkage and selection via the Lasso. *J. R. Stat. Soc. Ser. B (Methodological)*, **58**, 267–288.

Tsirigos,K.D. *et al.* (2015) The TOPCONS web server for consensus prediction of membrane protein topology and signal peptides. *Nucleic Acids Res.*, **43**, W401–W407.

Vanacek,P. *et al.* (2018) Exploration of enzyme diversity by integrating bioinformatics with expression analysis and biochemical characterization. *ACS Catal.*, **8**, 2402–2412.

Walsh,I. *et al.* (2012) ESpritz: accurate and fast prediction of protein disorder. *Bioinformatics*, **28**, 503–509.

Wilkinson,D.L. and Harrison,R.G. (1991) Predicting the solubility of recombinant proteins in *Escherichia coli*. *Biotechnology (N.Y.)*, **9**, 443–448

A.7 Paper VII

EnzymeMiner: automated mining of soluble enzymes with diverse structures, catalytic properties and stabilities

doi: 10.1093/nar/gkaa372

EnzymeMiner: automated mining of soluble enzymes with diverse structures, catalytic properties and stabilities

Jiri Hon[1,2,3,†], Simeon Borko[1,2,†], Jan Stourac[1,3], Zbynek Prokop[1,3], Jaroslav Zendulka[2], David Bednar[1,3], Tomas Martinek[2] and Jiri Damborsky[1,3,*]

[1]Loschmidt Laboratories, Department of Experimental Biology and Research Center for Toxic Compounds in the Environment RECETOX, Faculty of Science, Masaryk University, Brno, Czech Republic, [2]IT4Innovations Centre of Excellence, Faculty of Information Technology, Brno University of Technology, Bozetechova 2, Brno, Czech Republic and [3]International Clinical Research Center, St. Anne's University Hospital Brno, Brno, Czech Republic

Received March 08, 2020; Revised April 13, 2020; Editorial Decision April 27, 2020; Accepted April 29, 2020

ABSTRACT

Millions of protein sequences are being discovered at an incredible pace, representing an inexhaustible source of biocatalysts. Despite genomic databases growing exponentially, classical biochemical characterization techniques are time-demanding, cost-ineffective and low-throughput. Therefore, computational methods are being developed to explore the unmapped sequence space efficiently. Selection of putative enzymes for biochemical characterization based on rational and robust analysis of all available sequences remains an unsolved problem. To address this challenge, we have developed EnzymeMiner—a web server for automated screening and annotation of diverse family members that enables selection of hits for wet-lab experiments. EnzymeMiner prioritizes sequences that are more likely to preserve the catalytic activity and are heterologously expressible in a soluble form in *Escherichia coli*. The solubility prediction employs the in-house SoluProt predictor developed using machine learning. EnzymeMiner reduces the time devoted to data gathering, multi-step analysis, sequence prioritization and selection from days to hours. The successful use case for the haloalkane dehalogenase family is described in a comprehensive tutorial available on the EnzymeMiner web page. EnzymeMiner is a universal tool applicable to any enzyme family that provides an interactive and easy-to-use web interface freely available at https://loschmidt.chemi.muni.cz/enzymeminer/.

INTRODUCTION

There are currently >259 million non-redundant protein sequences in the NCBI nr database (release 2020-02-10) (1). Despite their enormous promise for biological and biotechnological discovery, experimental characterization has been performed on only a small fraction of the available sequences. Currently, there are about 560 000 protein sequences reliably curated in the UniProtKB/Swiss-Prot database (release 2020_01) (2).

The low ratio of characterized to uncharacterized sequences reflects the sharp contrast in time-demanding/low-throughput biochemical techniques versus fast/high-throughput next-generation sequencing technology. Although more efficient biochemical techniques employing miniaturization and automation have been developed (3–5), the most widely used experimental methods do not provide sufficient capacity for biochemical characterization of proteins spanning the ever-increasing sequence space. Therefore, computational methods are currently the only way to explore the immense protein diversity available among the millions of uncharacterized sequence entries.

Two different computational strategies are generally used for exploration of the unknown sequence space. The first strategy takes a novel uncharacterized sequence as input and predicts functional annotations. The method involves annotating the unknown input sequences by predicting protein domains (6), Enzyme Commission (EC) number (7) or Gene Ontology terms that are a subject of the initiative named the Critical Assessment of Functional Annotation (8). These methods are often universal and applicable to any protein sequence. However, they often lack specificity as the automatic annotation rules or statistical models need to be substantially general. A significant advantage of these methods is their seamless integration into available

[*]To whom correspondence should be addressed. Tel: +420 5 4949 3467; Email: jiri@chemi.muni.cz

[†]The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

databases. Submission of a query sequence to a database is sufficient, with no need for running computation- and memory-intensive bioinformatics pipelines locally. A model example of this approach is the automatic annotation workflow of the UniProtKB/TrEMBL database (2).

The second strategy takes a well-known characterized sequence as an input and applies a computational workflow, typically based on a homology search, to identify novel uncharacterized entries in genomic databases that are related to the input query sequence (5,9). The homology search is often followed by a filtration step, which checks the essential sequence properties, e.g. domain structure or presence of catalytic residues. The main advantage of these methods is the higher specificity of the analysis. A disadvantage is that it may be complicated to apply the developed workflow to protein families other than those for which it was designed. Moreover, these workflows typically require running complex bioinformatics pipelines and are usually not available through a web interface.

The fundamental unsolved problem is how to deal with the overwhelming number of sequence entries identified by these methods and select a small number of relevant hits for in-depth experimental characterization. For example, a database search for members of the haloalkane dehalogenase model family using the UniProt web interface yields 3598 sequences (UniProtKB release 2020_01). It is impossible to rationally select several tens of targets for experimental testing without additional bioinformatics analyses to help prioritize such a large pool of sequences.

To address the challenge of exploring the unmapped enzyme sequence space and rational selection of attractive targets, we have developed the EnzymeMiner web server. EnzymeMiner identifies novel enzyme family members, comprehensively annotates the targets and facilitates efficient prioritization and selection of representative hits for experimental characterization. To the best of our knowledge, there is currently no other tool available that allows such a comprehensive analysis in a single easy-to-run integrated workflow on the web.

MATERIALS AND METHODS

EnzymeMiner implements a three-step workflow: (i) homology search, (ii) essential residue based filtering and (iii) hits annotation (Figure 1). To execute these tasks, the server requires two different types of input information: (i) query sequences and (ii) essential residue templates. The query sequences serve as seeds for the initial homology search. The essential residue templates, defined as pairs of a protein sequence and a set of essential residues in that sequence, allow the server to prioritize hits that are more likely to display the enzyme function. Therefore, the essential residues may be the catalytic and ligand- or cofactor-binding residues that are indispensable for proper catalytic function. Each essential residue is defined by its name, position and a set of allowed amino acids for that position.

In the first *homology search step*, a query sequence is used as a query for a PSI-BLAST (10) two-iteration search in the NCBI nr database (1). If more than one query sequence is provided, a search is conducted for each sequence separately. Besides a minimum *E*-value threshold 10^{-20}, the PSI-BLAST hits must share a minimum of 25% global sequence identity with at least one of the query sequences. Artificial protein sequences, i.e. sequences described by the term artificial, synthetic construct, vector, vaccinia virus, plasmid, halotag or replicon, are removed. EnzymeMiner sorts the PSI-BLAST hits by *E*-value and passes a maximum of 10,000 best hits to the next steps in the workflow. The default parameters for the homology search step, as well as the other steps, can be modified using advanced options in the web server.

In the second *essential residue based filtering step*, the homology search hits are filtered using the essential residue templates. First, the hits are divided into template clusters. Each cluster contains all hits matching essential residues of a particular template. Essential residues are checked using global pairwise alignment with the template calculated by USEARCH (11). When multiple essential residue templates match, the hit is assigned to the template with the highest global sequence identity. Second, for each cluster, an initial multiple sequence alignment (MSA) is constructed using Clustal Omega (12). The MSA is used to revalidate the essential residues of identified hits by checking the corresponding column in the MSA. Sequences not matching essential residues of the template are removed from the cluster. Third, the MSA is constructed again for each template cluster and the essential residues are checked for the last time. The final set of identified sequences reported by EnzymeMiner contains all sequences left in the template clusters.

In the third *annotation step*, the identified sequences are annotated using several databases and predictors: (i) transmembrane regions are predicted by TMHMM (13), (ii) Pfam domains are predicted by InterProScan (14), (iii) source organism annotation is extracted from the NCBI Taxonomy (15) and the NCBI BioProject database (16), (iv) protein solubility is predicted by the in-house tool SoluProt for prediction of soluble protein expression in *Escherichia coli* and (v) sequence identities to queries, hits or other optional sequences are calculated by USEARCH (11). SoluProt is based on a random forest regression model that employs 36 sequence-based features (https://loschmidt.chemi.muni.cz/soluprot/). It has been shown to achieve an accuracy of 58%, specificity of 73% and sensitivity of 44% on a balanced independent test set of 3788 sequences (Hon et al., manuscript in preparation). Alternative solubility prediction tools are summarised in a recently published review (17). It is not advised to use the solubility score for other expression systems because it was trained solely on *E. coli* data. We expect further intensive development of protein solubility predictors in coming years and will ensure that the solubility score in the EnzymeMiner stays at the cutting-edge in terms of its accuracy and reproducibility.

The sequence space of the identified hits is visualized using representative sequence similarity networks (SSNs) generated at various clustering thresholds using MMseqs2 (18) and Cytoscape (19). SSNs provide a clean visual approach to identify clusters of highly similar sequences and rapidly spot sequence outliers. SSNs proved to facilitate identification of previously unexplored sequence and function space (20). The SSN generation method used in EnzymeMiner is inspired by the EFI-EST tool (21). The minimum align-

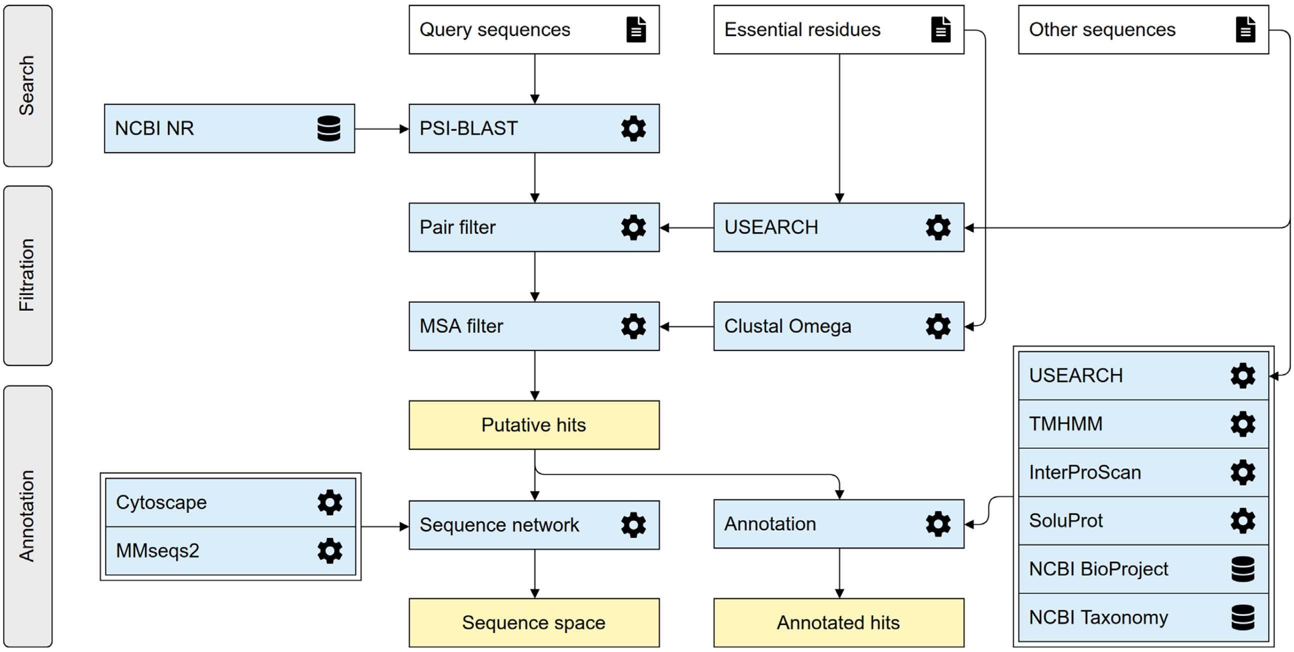


Figure 1. The EnzymeMiner workflow. The workflow consists of three distinct steps: (i) sequence homology search, (ii) filtration of functional sequences, and (iii) annotation of hits. These steps are executed consecutively and automatically. EnzymeMiner has only two required inputs: (i) query sequences, and (ii) essential residue templates. The *Other sequences* are optional inputs that allow EnzymeMiner to calculate the sequence identity between these sequences and all the hits. Input files are highlighted by a white background, tools and databases have a light blue background, outputs are highlighted by a yellow background.

ment score to include an edge between two representative sequences in an SSN is 40.

DESCRIPTION OF THE WEB SERVER

Job submission

New jobs can be submitted from the EnzymeMiner homepage. EnzymeMiner provides two conceptually different ways to define the input of the workflow: (i) using curated sequences from the UniProtKB/Swiss-Prot database and (ii) using custom sequences. We recommend the UniProtKB/Swiss-Prot option for users who do not have in-depth knowledge of the enzyme family. In contrast, the *Custom sequences* tab gives full control over the EnzymeMiner input—query sequences and essential residue templates are specified manually by the user. This is recommended for users who have good knowledge about the enzyme family and want to provide additional starting information to obtain refined results. The last option is a combination of both approaches, where Swiss-Prot sequences can be pre-selected first and then the input can be modified in the *Custom sequences* tab.

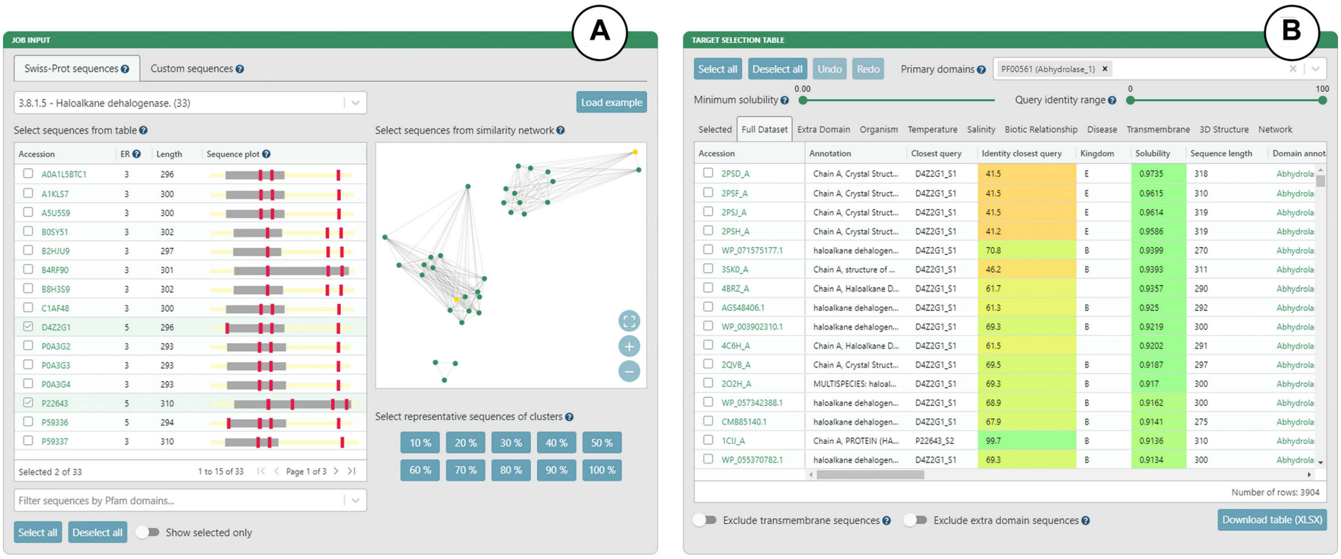
In the *Swiss-Prot sequences* tab (Figure 2A), sequences from the Swiss-Prot database can be queried by Enzyme Commission (EC) number. As a result, a table of all sequences annotated by the EC number and corresponding SSN is generated. The table has four columns: (i) sequence accessions hyperlinked to the UniProt database, (ii) number of essential residues, (iii) sequence length and (iv) sequence plot. The sequence plot summarizes two important features of the sequence – positions of essential residues and identified Pfam domains. The positions of essential residues are obtained from the Swiss-Prot database. The SSN visualizes the sequence space of all the sequences in the current EC group. Nodes represent Swiss-Prot sequences, whereas edge lengths are proportional to the pairwise sequence identities. Similar sequences are close to each other, whereas more distant sequences are not connected at all.

There are three strategies possible for selecting Swiss-Prot sequences as the EnzymeMiner query: (i) select a row from the sequence table, (ii) select a node in the SSN and (iii) select cluster representatives by defining a sequence identity threshold. The sequence identity threshold buttons select cluster representatives at the given percentage threshold. Using this feature, the user can automatically select a small set of sequences that cover the whole known sequence space of the current EC group. All selected Swiss-Prot sequences are used as a query in the homology search step and also as essential residue templates for the filtration step. To modify the selected sets of queries and essential residue templates, the user can switch to the *Custom sequences* tab and refine the selection manually.

EnzymeMiner results

The results page is organized into four sections: (i) *job information* box, (ii) *download results* box, (iii) *target selection table* and (iv) *sequence similarity network*.

In the *job information* box, the user can find the job ID, title, start time and status of the job. There is also a rerun button for rerunning the same analysis without the need for re-entering the same input. This feature is handy for periodically mining new sequences as the sequence databases



3.8.1.5 - Haloalkane dehalogenase. (33)

Select sequences from table

	Accession	ER	Length	Sequence plot
□	A0A1L5BTC1	3	296	
□	A1KLS7	3	300	
□	A5U5S9	3	300	
□	B0SY51	3	302	
□	B2HJU9	3	297	
□	B4RF90	3	301	
□	B8H3S9	3	302	
□	C1AF48	3	300	
☑	D4Z2G1	5	296	
□	P0A3G2	3	293	
□	P0A3G3	3	293	
□	P0A3G4	3	293	
☑	P22643	5	310	
□	P59336	5	294	
□	P59337	3	310	

Selected 2 of 33 1 to 15 of 33 |< < Page 1 of 3 > >|

Filter sequences by Pfam domains...

Select all Deselect all Show selected only

Primary domains: PF00561 (Abhydrolase_1)

Minimum solubility 0.00

Query identity range 0 100

	Accession	Annotation	Closest query	Identity closest query	Kingdom	Solubility	Sequence length	Domain annot
□	2PSD_A	Chain A, Crystal Struct...	D4Z2G1_S1	41.5	E	0.9735	318	Abhydrola
□	2PSF_A	Chain A, Crystal Struct...	D4Z2G1_S1	41.5	E	0.9615	310	Abhydrola
□	2PSJ_A	Chain A, Crystal Struct...	D4Z2G1_S1	41.5	E	0.9614	319	Abhydrola
□	2PSH_A	Chain A, Crystal Struct...	D4Z2G1_S1	41.2	E	0.9586	319	Abhydrola
□	WP_071575177.1	haloalkane dehalogen...	D4Z2G1_S1	70.8	B	0.9399	270	Abhydrola
□	3SK0_A	Chain A, structure of ...	D4Z2G1_S1	46.2	B	0.9393	311	Abhydrola
□	4BRZ_A	Chain A, Haloalkane D...	D4Z2G1_S1	61.7		0.9357	290	Abhydrola
□	AGS48406.1	haloalkane dehalogen...	D4Z2G1_S1	61.3	B	0.925	292	Abhydrola
□	WP_003902310.1	haloalkane dehalogen...	D4Z2G1_S1	69.3	B	0.9219	300	Abhydrola
□	4C6H_A	Chain A, Haloalkane D...	D4Z2G1_S1	61.5		0.9202	291	Abhydrola
□	2QVB_A	Chain A, Crystal Struct...	D4Z2G1_S1	69.5	B	0.9187	297	Abhydrola
□	2O2H_A	MULTISPECIES: haloal...	D4Z2G1_S1	69.3	B	0.917	300	Abhydrola
□	WP_057342388.1	haloalkane dehalogen...	D4Z2G1_S1	68.9	B	0.9162	300	Abhydrola
□	CMB85140.1	haloalkane dehalogen...	D4Z2G1_S1	67.9	B	0.9141	275	Abhydrola
□	1CIJ_A	Chain A, PROTEIN (HA...	P22643_S2	99.7	B	0.9136	310	Abhydrola
□	WP_055370782.1	haloalkane dehalogen...	D4Z2G1_S1	69.3	B	0.9134	300	Abhydrola

Number of rows: 3904

Exclude transmembrane sequences Exclude extra domain sequences

Figure 2. The EnzymeMiner graphical user interface showing example inputs and results for the haloalkane dehalogenase family (EC 3.8.1.5). (**A**) Inputs based on curated sequences from the UniProtKB/Swiss-Prot database. The input sequences can be selected using: (i) the sequence table, (ii) the SSN or (iii) the sequence identity threshold. (**B**) Target selection table. The table is organized into eleven sheets that summarize the results from different perspectives. The table can be filtered using solubility and identity sliders, and transmembrane and extra domain exclusion switches.

grow. For example, there are hundreds of new hits for the haloalkane dehalogenase family every year. In the *download results* box, the user can download the results table in XLSX format or tab-separated text format. A ZIP archive containing all output files from the EnzymeMiner workflow can also be downloaded.

The *target selection table* is the most important component of the EnzymeMiner results (Figure 2B). It presents all the putative enzyme sequences identified by EnzymeMiner and helps to select targets for experimental characterization. The table is organized into eleven sheets summarizing the results from different perspectives. (i) The *Selected* sheet shows all the sequences selected from individual sheets. It contains an extra column to track the argument used for the selection. By default, it is prefilled by the name of the sheet from which the sequence was selected, but it can be freely changed. (ii) The *Full Dataset* sheet shows all identified sequences. (iii) The *Extra domain* sheet shows sequences with extra Pfam domains found in the sequence but not listed in the *Primary domains* selection box. (iv) The *Organism* sheet shows sequences with known source organisms. (v) The *Temperature* sheet shows sequences from organisms having extreme optimum temperature annotation in the NCBI BioProject database, including sequences from thermophilic or cryophilic organisms. (vi) The *Salinity* sheet shows sequences from organisms having extreme salinity annotation in the NCBI BioProject database. (vii) The *Biotic Relationship* sheet shows sequences from organisms having biotic relationship annotation in the NCBI BioProject database. (viii) The *Disease* sheet shows sequences from organisms having disease annotation in the NCBI BioProject database. (ix) The *Transmembrane* sheet shows sequences with transmembrane regions predicted by the TMHMM tool. (x) The *3D Structure* sheet shows sequences with an available 3D structure in the Protein Data Bank (22). (xi) The *Network* sheet shows sequences clustered into a selected sequence similarity network node.

There are four options for filtering the identified sequences displayed in the target selection table. The first option is the minimum solubility slider. Sequences with lower predicted solubility will be hidden. We recommend setting the solubility threshold to >0.5 to increase the probability of finding soluble protein expression in *E. coli*. We do not recommend to set the solubility threshold too high because of possible trade-off between enzyme solubility and activity (23). The second option is the identity range bar. Only sequences with identity to query sequences in the specified range will be visible. The third option is to exclude transmembrane proteins. We recommend removing these sequences as they are usually difficult to produce and tend to have lower predicted solubility. The fourth option is to exclude proteins with an extra domain. Extra domains are defined as domains found in the sequence but not listed in the *Primary domains* selection box. We recommend avoiding sequences with extra domains, but these sequences may also show interesting and unusual biological properties. The selection table can be sorted by clicking on a column header. Holding 'Shift' while clicking on the column headers allows sorting by multiple columns.

The SSN visualizes the sequence space of all identified sequences. Both clusters of similar sequences and sequence outliers can be easily identified. As there might be thousands of sequences, the sequences are clustered at the identity threshold and only an SSN of the representative sequences is shown for performance reasons. Sequences having greater sequence identity are consolidated into a single metanode. Edges indicate high sequence identity between representative sequences of the connected metanodes. Clicking on a metanode displays the *Network* sheet showing which sequences are represented by a particular metanode. The SSN can be downloaded as a Cytoscape session file for further analysis and custom visualization. Networks clustered at different identities are available. The numbers of nodes and edges are indicated for each identity threshold. The SSN is interactively linked to the target selection table. All nodes representing selected sequences are automatically highlighted in the SSN.

Target selection

The target selection table and SSN facilitate the selection of a diverse set of soluble putative enzyme sequences for experimental validation. First, we recommend setting the maximum sequence identity to queries to 90%. This will remove all hits that are very similar to already known proteins. Second, we recommend selecting a few sequences from individual sheets to cover different phyla from the domains Archea, Bacteria and Eukarya. The most exciting enzymes might be from extremophilic organisms. Third, the SSN can be used to check that the selection covers all sequence clusters. Fourth, users can select sequences from all subfamilies of the enzyme family of interest. The members of different subfamilies can be easily recognized by the *Closest query* or *Closest known* column in the selection table (note: requires representative sequences of subfamilies as job input). Fifth, the available filtering options can be used to (i) prioritize sequences with the highest predicted solubility, (ii) prioritize sequences with known tertiary structures, (iii) eliminate proteins with predicted transmembrane regions and (iv) eliminate sequences with extra domains.

EXPERIMENTAL VALIDATION OF THE EnzymeMiner WORKFLOW

The EnzymeMiner workflow has been thoroughly experimentally validated using the model enzymes haloalkane dehalogenases (5). The sequence-based search identified 658 putative dehalogenases. The subsequent analysis prioritized and selected 20 candidate genes for exploration of their protein structural and functional diversity. The selected enzymes originated from genetically unrelated Bacteria, Eukarya and, for the first time, also Archaea and showed novel catalytic properties and stabilities. The workflow helped to identify novel haloalkane dehalogenases, including (i) the most catalytically efficient enzyme ($k_{\mathrm{cat}}/\mathrm{K}_{0.5} = 96.8\ \mathrm{mM}^{-1}\ \mathrm{s}^{-1}$), (ii) the most thermostable enzyme showing a melting temperature of 71°C, (iii) three different cold-adapted enzymes active at near to 0°C, (iv) highly enantioselective enzymes, (v) enzymes with a wide range of optimal operational temperature from 20 to 70°C and an unusually broad pH range from 5.7–10 and (vi) biocatalysts degrading the warfare chemical yperite and various environmental pollutants. The sequence mining, annotation, and visualization steps from the workflow published by Vanacek and coworkers (5) were fully automated in the EnzymeMiner web server. The successful use case for the haloalkane dehalogenase family is described in an easy-to-follow tutorial available on the EnzymeMiner web page. Additional extensive validation of the fully automated version of EnzymeMiner, experimentally testing the properties of another 45 genes of the haloalkane dehalogenases, is currently ongoing in our laboratory.

CONCLUSIONS AND OUTLOOK

The EnzymeMiner web server identifies putative members of enzyme families and facilitates their prioritization and well-informed manual selection for experimental characterization to reveal novel biocatalysts. Such a task is difficult using the web interfaces of the available protein databases, e.g. UniProtKB/TrEMBL and NCBI Protein, since additional analyses are often required. The major advantage of EnzymeMiner over existing protein sources is the flexibility of input and concise annotation-rich interactive presentation of results. The user can input custom queries and a custom description of essential residues to focus the search on specific protein families or subfamilies. The output of EnzymeMiner is an interactive selection table containing the annotated sequences divided into sheets based on various criteria. The table helps to select a diverse set of sequences for experimental characterization. Two key prioritization criteria are (i) the predicted solubility score, which can be used to prioritize the identified sequences and increase the chance of finding enzymes with soluble protein expression, and (ii) the sequence identity to query sequences complemented with an interactive SSN displayed directly on the web, which can be used to find diverse sequences. Additionally, source organism and domain annotations help to select sequences with diverse properties. EnzymeMiner is a universal tool applicable to any enzyme family. It reduces the time needed for data gathering, multi-step analysis and sequence prioritization from days to hours. All the EnzymeMiner features are implemented directly on the web server and no external tools are required. The web server was optimized for modern browsers including Chrome, Firefox and Safari. An EnzymeMiner job can take a few hours or days to compute, depending on the current load of the server. In the next EnzymeMiner version, we plan three major improvements. First, we will implement automated tertiary structure prediction based on homology modeling and threading for all identified sequences. The structural predictions will allow subsequent analysis of active site pockets/cavities and access tunnels. Structural features will significantly enrich the set of annotations and help to identify additional attractive targets for experimental characterization. Second, we will implement automated periodical mining. When enabled, EnzymeMiner will rerun the analysis periodically and inform the user about novel sequences found since the last search. Finally, we will implement a wizard for automated selection of hits based on input criteria provided by a user.

ACKNOWLEDGEMENTS

We thank the participants of the 1st Hands-on Computational Enzyme Design Course (Brno, Czech Republic) for giving valuable feedback on the EnzymeMiner user interface. Their comments inspired us to make the sequence similarity network visualization more interactive.

FUNDING

Czech Ministry of Education [857560, 02.1.01/0.0/0.0/18_046/0015975, CZ.02.1.01/0.0/0.0/16_026/0008451, CZ.02.1.01/0.0/0.0/16_019/0000868, LQ1602]; European Commission [720776, 814418]; AI Methods for Cybersecurity and Control Systems project of the Brno University of Technology [FIT-S-20-6293]; Computational resources were supplied by the project 'e-Infrastruktura CZ' [e-INFRA LM2018140] provided within the program Projects of Large Research, Development and Innovations Infrastructures, and by the ELIXIR-CZ project [LM2015047], part of the international ELIXIR infrastructure. Funding for open access charge: Czech Ministry of Education.

Conflict of interest statement. None declared.

REFERENCES

1. Sayers,E.W., Agarwala,R., Bolton,E.E., Brister,J.R., Canese,K., Clark,K., Connor,R., Fiorini,N., Funk,K., Hefferon,T. *et al.* (2019) Database resources of the national center for biotechnology information. *Nucleic Acids Res.*, **47**, D23–D28.
2. UniProt Consortium (2019) UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.*, **47**, D506–D515.
3. Colin,P.-Y., Kintses,B., Gielen,F., Miton,C.M., Fischer,G., Mohamed,M.F., Hyvönen,M., Morgavi,D.P., Janssen,D.B. and Hollfelder,F. (2015) Ultrahigh-throughput discovery of promiscuous enzymes by picodroplet functional metagenomics. *Nat. Commun.*, **6**, 1–12.
4. Beneyton,T., Thomas,S., Griffiths,A.D., Nicaud,J.-M., Drevelle,A. and Rossignol,T. (2017) Droplet-based microfluidic high-throughput screening of heterologous enzymes secreted by the yeast Yarrowia lipolytica. *Microb. Cell Fact.*, **16**, 18.
5. Vanacek,P., Sebestova,E., Babkova,P., Bidmanova,S., Daniel,L., Dvorak,P., Stepankova,V., Chaloupkova,R., Brezovsky,J., Prokop,Z. *et al.* (2018) Exploration of enzyme diversity by integrating bioinformatics with expression analysis and biochemical characterization. *ACS Catal.*, **8**, 2402–2412.
6. El-Gebali,S., Mistry,J., Bateman,A., Eddy,S.R., Luciani,A., Potter,S.C., Qureshi,M., Richardson,L.J., Salazar,G.A., Smart,A. *et al.* (2019) The Pfam protein families database in 2019. *Nucleic Acids Res.*, **47**, D427–D432.
7. Li,Y., Wang,S., Umarov,R., Xie,B., Fan,M., Li,L. and Gao,X. (2018) DEEPre: sequence-based enzyme EC number prediction by deep learning. *Bioinformatics*, **34**, 760–769.
8. Zhou,N., Jiang,Y., Bergquist,T.R., Lee,A.J., Kacsoh,B.Z., Crocker,A.W., Lewis,K.A., Georghiou,G., Nguyen,H.N., Hamid,M.N. *et al.* (2019) The CAFA challenge reports improved protein function prediction and new functional annotations for hundreds of genes through experimental screens. *Genome Biol.*, **20**, 244.
9. Mak,W.S., Tran,S., Marcheschi,R., Bertolani,S., Thompson,J., Baker,D., Liao,J.C. and Siegel,J.B. (2015) Integrative genomic mining for enzyme function to enable engineering of a non-natural biosynthetic pathway. *Nat. Commun.*, **6**, 1–10.
10. Altschul,S. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
11. Edgar,R.C. (2010) Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*, **26**, 2460–2461.
12. Sievers,F., Wilm,A., Dineen,D., Gibson,T.J., Karplus,K., Li,W., Lopez,R., McWilliam,H., Remmert,M., Söding,J. *et al.* (2011) Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.*, **7**, 539.
13. Krogh,A., Larsson,B., von Heijne,G. and Sonnhammer,E.L. (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol.*, **305**, 567–580.
14. Quevillon,E., Silventoinen,V., Pillai,S., Harte,N., Mulder,N., Apweiler,R. and Lopez,R. (2005) InterProScan: protein domains identifier. *Nucleic Acids Res.*, **33**, W116–W120.
15. Federhen,S. (2012) The NCBI Taxonomy database. *Nucleic Acids Res.*, **40**, D136–D143.
16. Barrett,T., Clark,K., Gevorgyan,R., Gorelenkov,V., Gribov,E., Karsch-Mizrachi,I., Kimelman,M., Pruitt,K.D., Resenchuk,S., Tatusova,T. *et al.* (2012) BioProject and BioSample databases at NCBI: facilitating capture and organization of metadata. *Nucleic Acids Res.*, **40**, D57–D63.
17. Musil,M., Konegger,H., Hon,J., Bednar,D. and Damborsky,J. (2019) Computational design of Stable and Soluble Biocatalysts. *ACS Catal.*, **9**, 1033–1054.
18. Steinegger,M. and Söding,J. (2017) MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat. Biotechnol.*, **35**, 1026–1028.
19. Shannon,P. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.*, **13**, 2498–2504.
20. Copp,J.N., Akiva,E., Babbitt,P.C. and Tokuriki,N. (2018) Revealing unexplored sequence-function space using sequence similarity networks. *Biochemistry*, **57**, 4651–4662.
21. Gerlt,J.A., Bouvier,J.T., Davidson,D.B., Imker,H.J., Sadkhin,B., Slater,D.R. and Whalen,K.L. (2015) Enzyme Function Initiative-Enzyme Similarity Tool (EFI-EST): A web tool for generating protein sequence similarity networks. *Biochim. Biophys. Acta (BBA) - Proteins Proteomics*, **1854**, 1019–1037.
22. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.
23. Klesmith,J.R., Bacik,J.-P., Wrenbeck,E.E., Michalczyk,R. and Whitehead,T.A. (2017) Trade-offs between enzyme fitness and solubility illuminated by deep mutational scanning. *Proc. Natl Acad. Sci. U.S.A.*, **114**, 2265–2270.

A.8 Paper VIII

HotSpot Wizard 3.0: web server for automated design of mutations and smart libraries based on sequence input information

W356–W362 *Nucleic Acids Research, 2018, Vol. 46, Web Server issue* *Published online 23 May 2018*
doi: 10.1093/nar/gky417

HotSpot Wizard 3.0: web server for automated design of mutations and smart libraries based on sequence input information

Lenka Sumbalova[1,2], Jan Stourac[1,3], Tomas Martinek[2], David Bednar[1,3,*] and Jiri Damborsky[1,3,*]

[1]Loschmidt Laboratories, Department of Experimental Biology, Masaryk University, 62500 Brno, Czech Republic, [2]IT4Innovations Centre of Excellence, Faculty of Information Technology, Brno University of Technology, Bozetechova 2, 61266 Brno, Czech Republic and [3]International Centre for Clinical Research, St. Anne's University Hospital Brno, 65691 Brno, Czech Republic

Received February 04, 2018; Revised April 20, 2018; Editorial Decision May 02, 2018; Accepted May 07, 2018

ABSTRACT

HotSpot Wizard is a web server used for the automated identification of hotspots in semi-rational protein design to give improved protein stability, catalytic activity, substrate specificity and enantioselectivity. Since there are three orders of magnitude fewer protein structures than sequences in bioinformatic databases, the major limitation to the usability of previous versions was the requirement for the protein structure to be a compulsory input for the calculation. HotSpot Wizard 3.0 now accepts the protein sequence as input data. The protein structure for the query sequence is obtained either from eight repositories of homology models or is modeled using Modeller and I-Tasser. The quality of the models is then evaluated using three quality assessment tools—WHAT_CHECK, PROCHECK and MolProbity. During follow-up analyses, the system automatically warns the users whenever they attempt to redesign poorly predicted parts of their homology models. The second main limitation of HotSpot Wizard's predictions is that it identifies suitable positions for mutagenesis, but does not provide any reliable advice on particular substitutions. A new module for the estimation of thermodynamic stabilities using the Rosetta and FoldX suites has been introduced which prevents destabilizing mutations among pre-selected variants entering experimental testing. HotSpot Wizard is freely available at http://loschmidt.chemi.muni.cz/hotspotwizard.

INTRODUCTION

Proteins are macromolecules with many biological functions. Apart from their irreplaceable role in all living organisms, they are also widely used in many fields, including medicine (1), enzymology (2), synthetic biology (3) and material science (4). Naturally occurring proteins often do not meet the specifications for practical applications. Therefore, protein engineers modify sequences to obtain enhanced properties or completely new functions. Directed evolution, which has been an extremely successful protein engineering technology, does not require a molecular understanding of the impact of mutation on the protein structure (5). Modified proteins are generated in iterative rounds of mutation and screening or selection of the best hits that possess the required property (6). The obvious disadvantage to this method is that only a tiny fraction of all protein variants contain the desired property. Analysis of libraries containing millions of mutants is costly and time-consuming. Semi-rational protein engineering is an approach that implements *in silico* identification of important regions of the protein so that mutagenesis is better located, resulting in smaller high-quality libraries (7). The key step to semi-rational protein engineering is the selection of hotspot residues whose mutations will bring the largest improvement to the target protein properties (8).

HotSpot Wizard 2.0 (9) is an interactive web server used for the identification of hotspots in proteins by automated multi-step calculation and a comprehensive presentation of results. The tool makes protein design accessible to researchers with no prior knowledge of bioinformatics. After entering an input protein structure, 19 prediction tools and 3 databases are used for protein annotation. HotSpot Wizard then provides four different strategies for selecting hotspots: (i) functional hotspots corresponding to highly mutable residues located in the active site

[*]To whom correspondence should be addressed. Tel: +420 5 4949 3467; Fax: +420 5 4949 6302; Email: jiri@chemi.muni.cz
[*]Correspondence may also be addressed to David Bednar. Email: 222755@mail.muni.cz

pocket or access tunnels, (ii) stability hotspots corresponding to flexible residues, (iii) stability hotspots from back-to-consensus analysis and (iv) correlated hotspots corresponding to pairs of co-evolving residues. The users can design a smart library based on naturally accepted substitutions from phylogenetic analysis. HotSpot Wizard 2.0 (9) has been used for over 10 000 protein structures by more than 1000 unique users since its release. For example, HotSpot Wizard has been used for the design of smart libraries of oxyhaemoglobin protein (10), for analysis leading to thermostabilization of a xylanase (11) and for identification of hotspots in a mutagenesis study of the transcription factor DREB1A (12). Previous implementations of HotSpot Wizard had two major drawbacks: (i) a requirement for the tertiary structure as essential input information and (ii) identification of positions for mutagenesis without quantification of the effects of individual substitutions on protein stability. HotSpot Wizard 3.0 shows dramatically enhanced usability by overcoming both these key limitations.

There are about 135 000 protein structures available in the RCSB Protein Data Bank (13), but there are more than 98 000 000 known protein sequences (14). Usage of HotSpot Wizard 2.0 is limited to the proteins with an available 3D structure. A solution to this problem is the prediction of the protein structure from its sequence by comparative (homology) modeling or threading (15). Homology modeling is based on the fact that members of a protein family with similar sequences also have similar tertiary structures (16,17). In HotSpot Wizard 3.0, it is possible to enter a sequence for a protein and have its tertiary structure retrieved from the repositories of models or constructed *ad hoc*. As the quality of the protein structure is critical for further structure analyses carried out by HotSpot Wizard, a robust quality assessment of the protein structure is provided using three well-established tools. The current implementation of our web server predicts hot-spots for mutagenesis and designs smart libraries based on phylogeny, but does not provide any quantitative analysis of individual substitutions, which is important, for example, in studies analyzing structure–function relationships. Moreover, screening or selection for multiple mutations at several different positions can still be time-consuming and so pre-selection of the most appropriate mutations is desirable. To help our users rationally decrease the number of variants for experimental testing, protein stability prediction has been introduced to discard potentially destabilizing mutations.

MATERIALS AND METHODS

Searches of structural databases and model depositories

The overall workflow of HotSpot Wizard 3.0 is outlined in Figure 1. When a protein sequence is used as an input, HotSpot Wizard: (i) searches experimentally determined structures, (ii) searches computationally modeled structures and (iii) constructs a homology model. The first step in this workflow is searching the RCSB Protein Data Bank (13). In this phase, only protein structures with a 100% sequence identity match (or part of the sequence matching the input with 100% sequence identity) are provided as a starting structure for the analysis. If no such structure is found, the Protein Model Portal (18) is searched. The Protein Model Portal collates models of protein structures from eight different resources: Center for Structures of Membrane Proteins, CSMP (19), Joint Center for Structural Genomics, JCSG (20), Midwest Center for Structural Genomics, MCSG (21), Northeast Structural Genomics Consortium, NESG (22), New York SGX Research Center for Structural Genomics, NYSGXRC (23), Joint Center for Molecular Modeling, JCMM (24), ModBase (25) and SWISS-MODEL Repository (26). HotSpot Wizard queries the Protein Model Portal and then lists all available hits. After selection of one of these models, the structure is downloaded directly to Hotspot Wizard from the repository.

Homology modeling

Whenever a homology model is not found or the user is not satisfied with the quality of the models available in public depositories, HotSpot Wizard carries out the homology modeling during the phase 1 (Figure 1). There is a wide range of homology modeling tools available. Twelve tools were initially considered for our workflow: SWISS-MODEL (27), Rosetta (28), Robetta (29), PHYRE2 (30), Pcons (31), Modeller (32), I-Tasser (33), IntFold (34), IMP (35), HHPred (36), RaptorX (37) and Sparks-X (38). These tools were analyzed for their availability as well as performance using Continuous Automated Model Evaluation, CAMEO (18) and Critical Assessment of Protein Structure Prediction, CASP (39). These community-wide comparisons evaluate structure predictions with available experimental data. Based on results from CASP and CAMEO, six tools were selected for further consideration, installed locally and tested (Modeller, Sparks-X, RaptorX, Rosetta, I-Tasser and SWISS-MODEL). RaptorX is very accurate with good coverage (i.e. percentage of submitted models, which could be successfully modeled), but it uses the less accurate Modeller for comparative modeling in its standalone version. Sparks-X is very fast with good coverage, but the version available for download does not provide modeling, only template identification. I-Tasser is the slowest of all the tools considered, but it is very accurate and is ranked the best by CASP. Rosetta has good accuracy and coverage, but it requires a template protein and an alignment as an input defined by user. SWISS-MODEL is fast with good coverage, but it is not available as a standalone version. Modeller is one of the fastest and the most robust tools with reasonable accuracy for modeling cases with good templates. We selected two tools for implementation with HotSpot Wizard: (i) I-Tasser, which is ranked the most accurate of all the tools considered, but also very slow (∼3 days for an average-sized protein) and (ii) Modeller, which is less accurate, but very fast (∼5 min for an average-sized protein). Both tools can be run in a fully automatic mode, or the template protein and/or the pairwise alignment can be entered as an input information.

Quality assessment of the model

It is essential to assess the quality of the homology model prior to its further use for identification of hotspots or for the design of libraries. It is important to identify low quality models and the parts of the protein structure which were

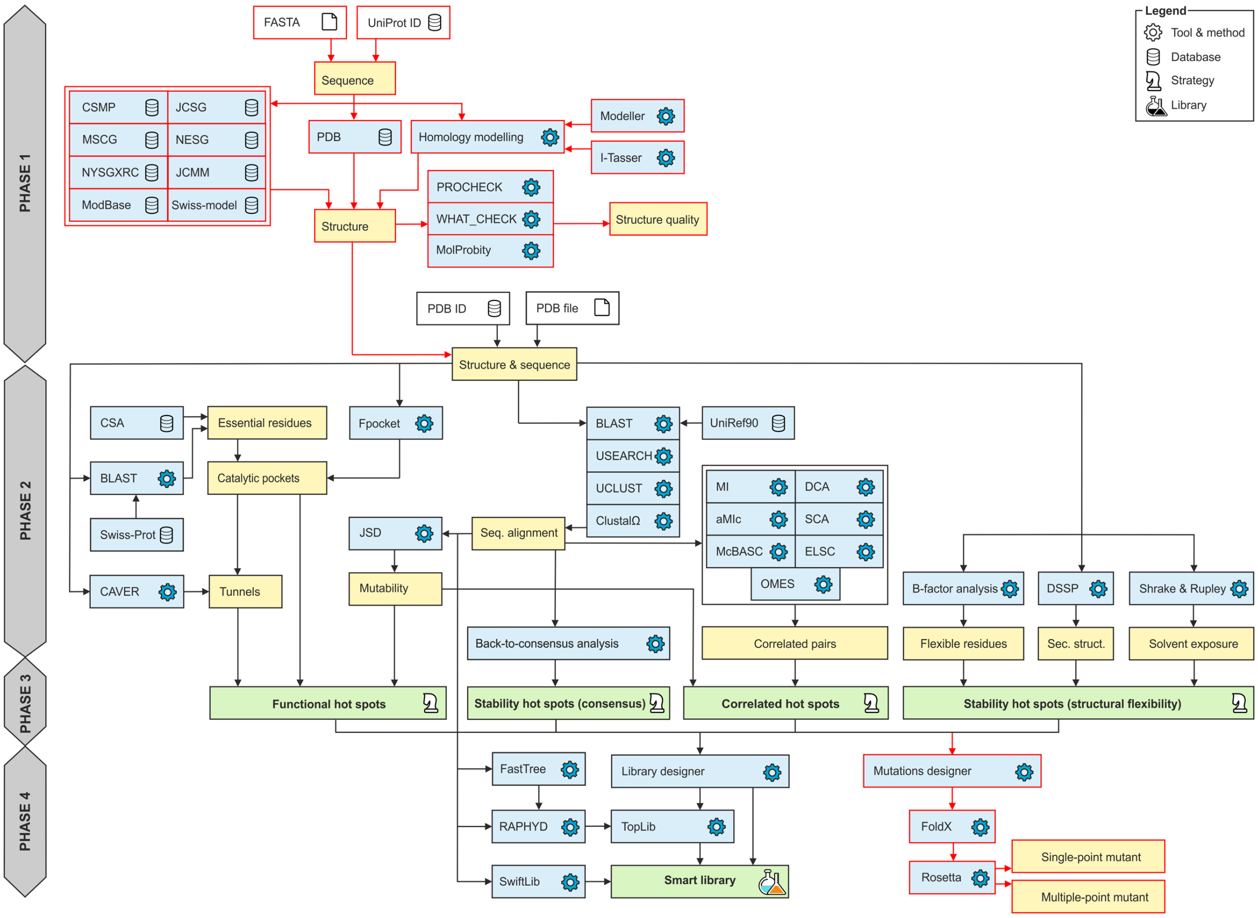


Figure 1. Workflow diagram of HotSpot Wizard 3.0. The workflow consists of four phases: (1) construction of a model of a structure, (2) annotation of a protein, (3) identification of mutagenesis hot spots and (4) design of mutations and a smart library. Phase 1 is applied only when a sequence is submitted as the input information. The new modules in version 3.0 are highlighted in red.

not modeled well. The results of today's modeling tools are far from perfect due to many difficulties with accurate protein structure prediction. Quality assessment is therefore an essential part of the phase 1 of the HotSpot Wizard workflow (Figure 1). Several quality assessment tools were considered and three of them, providing diverse quality metrics, were implemented. PROCHECK (40) is used for analysis of protein backbone torsion angles using Ramachandran diagrams and identification of the outliers from the allowed values. MolProbity (41) provides several parameters representing the quality of the whole structure as well as individual residues (number of poor rotamers, Ramachandran outliers, favored Ramachandran conformations, bad bonds and bad angles in the protein). WHAT_CHECK (42) generates a detailed report about structure quality (checks on secondary structure, coordinate problems, unexpected atoms, B-factor, occupancy checks, nomenclature related problems, geometric checks, torsion-related checks, bump checks, packing, accessibility, threading, water, ion and hydrogen bond-related checks).

Mutation design based on thermodynamic stability

Mutation design is part of the phase 4 of the HotSpot Wizard computation (Figure 1). Force field calculations are used for quantifying the change in protein thermodynamic stability after mutation. Rosetta (43) is used to evaluate ΔΔG between the wild-type and the mutant structures. Either single-point or multiple-point mutants can be evaluated. If the single-point mutations are pre-selected, multiple mutant structures are evaluated according to the user's selected positions and intended amino acid substitutions. The user can also select several mutations in a single round and calculate the energy of combined multiple-point mutants. For stability evaluation, FoldX (44) is first used for repairing protein structure by filling in the missing atoms and patching the structure. Then, minimalization of the structure using Rosetta is carried out using default settings. After that, a Rosetta stability calculation according to protocol 3 (45) is carried out, which results in the prediction of ΔΔG value for each mutation.

DESCRIPTION OF THE WEB SERVER

Sequence input and homology modeling

Initially, the user selects one of two types of input data: a structure or a sequence (Figure 2A). If a sequence is selected, there are three types of input. The user can either manually enter the protein sequence, specify the UniProt ID or upload the FASTA file. After entering the sequence, the user is provided with the results from searching the Protein Data Bank or the Protein Model Portal. This result is displayed in the form of a table (Figure 2B). In the case of the Protein Data Bank results, PDB ID, resolution and the link to the Protein Data Bank are provided. The user can then pick one of the proteins and continue with the HotSpot Wizard workflow. In the case of the results from the Protein Model Portal model provider, following information is listed: (i) used template, (ii) sequence identity with a template, (iii) range of the alignment, (iv) coverage and (v) reliability of the model. Links to a model in the Protein Model Portal and the template structure in the Protein Data Bank are provided in the table. Coverage and reliability of the models are represented by a color ranging from green to red (Figure 2C). If the user selects a model with unsatisfactory coverage (<80%) or insufficient reliability (low reliability value), a warning is displayed. When a protein model is selected which cannot be downloaded automatically, the user is asked to download it manually and then upload it as a structure for further analysis. The user can then select one of the models provided and continue with the HotSpot Wizard workflow or, if none of the models is satisfactory, carry out homology modeling and construct their own model. If the user carries out homology modeling, several parameters must be set first (Figure 2D). The user can select between Modeller, which is faster but less accurate, or I-Tasser, which is more accurate but slow. The second important parameter that must be specified prior to calculation is either automatic or manual identification of the template structure and alignment. The template can be provided either by entering the PDB ID or by uploading a PDB file. In the case of the user entering the alignment, pairwise alignment of the template and an input sequence in FASTA format must be provided. The process of hotspot identification can then begin after all these essential inputs have been defined.

Quality assessment of the model

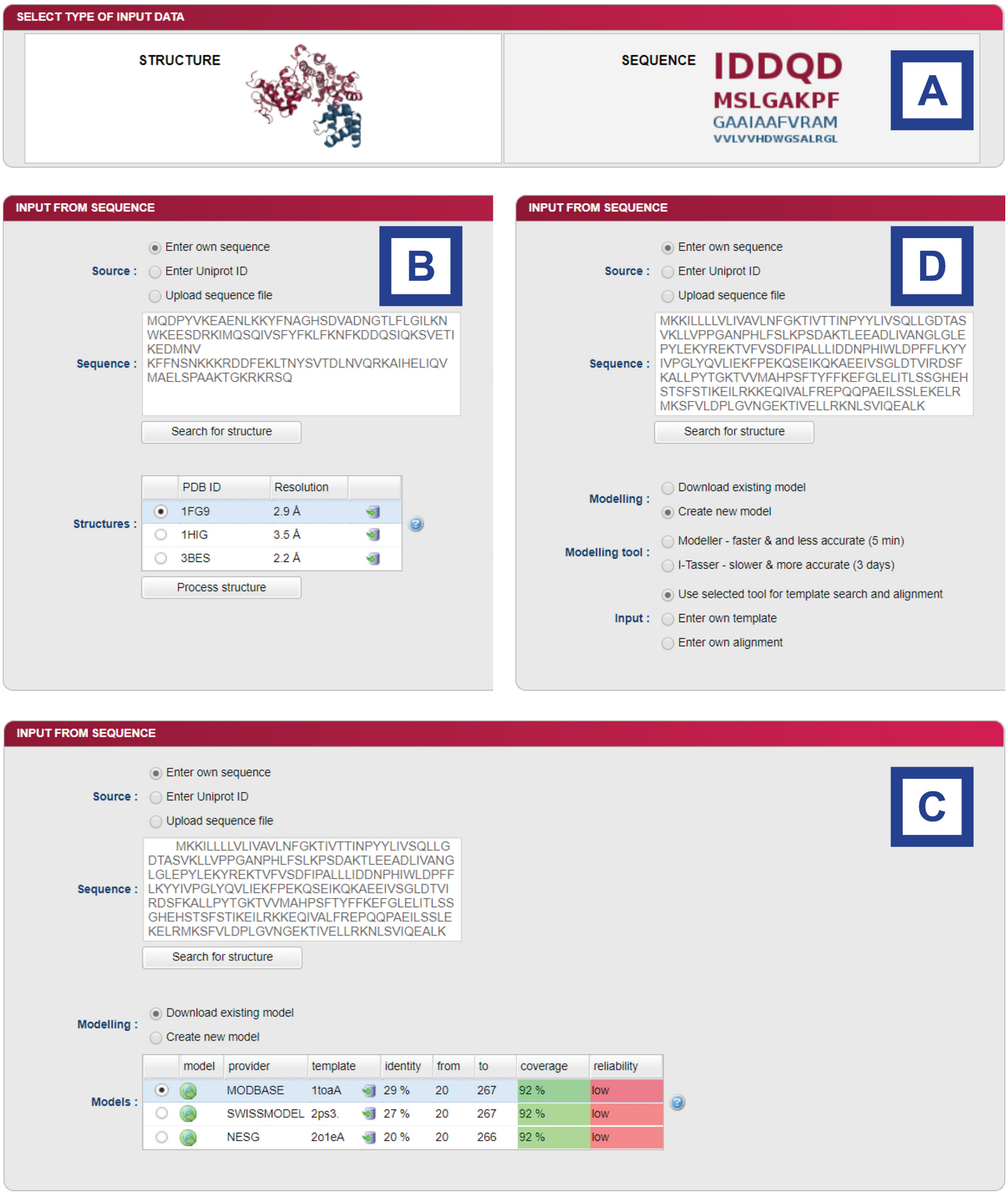
Results of the quality assessment are shown in separate windows consisting of three tabs containing various quality assessment analyses. The first tab shows the MolProbity overall quality assessment table (Supplementary Figure S1A). In this table, the number and percentage of poor rotamers, Ramachandran outliers, favored Ramachandran conformers, bad bonds and bad angles are shown. Colored highlights are used to distinguish between good and unsatisfactory models. The second tab shows the MolProbity quality assessment results for each residue, displayed in the form of plots (Supplementary Figure S1B). A plot of MolProbity Ramachandran scores and MolProbity rotamer scores is given. In the last tab, there is a Ramachandran plot for the protein created by PROCHECK with outlier residues highlighted (Supplementary Figure S1C). The contents of all these tabs can be downloaded in PDF format together with a full quality assessment report created by WHAT_CHECK.

Mutations design based on stability

The stability changes introduced by specific mutations can be accessed through a newly introduced Mutations design module (Supplementary Figure S2A). There are three tabs in the Mutation design window—the first for definition of single-point mutants, the second for multiple-point mutants and the third summarizing the status of submitted jobs. In the case of single-point mutations, the user can select particular amino acids for each of the selected hotspots. The amino acid residues for mutagenesis can be selected based on: (i) amino acid frequency, (ii) mutational landscape, (iii) physico-chemical properties or (iv) user selection (Supplementary Figure S2B). After selection of the mutations, the stability of each single-point mutation is evaluated by the Rosetta software suite. The results are shown in the table—stabilizing mutations are highlighted in green, destabilizing mutations are highlighted in red (Supplementary Figure S2C). There are two options for setting multiple-point mutants. Either a particular amino acid can be selected for each position in the multiple-point tab or the results table from a previous single-point calculation can be used for recombination with the most promising substitutions. In both cases, only a single substitution for each position can be selected (Supplementary Figure S2D). After the calculation is finished, Hotspot Wizard reports the overall stability change as well as the decomposition of energy terms, both of which provide excellent assistance for mutagenesis experiments (Supplementary Figure S2E). The stability prediction can be downloaded in CSV format with the sequence of designed mutants being provided in FASTA format. These reports can also be generated in PDF or HTML formats. The third tab shows a table with the history of previously evaluated stabilities for the job. For each calculation, the job id, date and time of computation, status of the job (failed or finished), mutation type (single-point or multiple-point), selected positions and mutations are shown (Supplementary Figure S2F). The results page from any previous calculations can be revisited at any time.

EXPERIMENTAL VALIDATION

We have carried out validation of individual steps of the workflow as well as thoroughly tested the final version of the web server. The homology modeling tools were selected for implementation based on the results of CAMEO comparison (Supplementary Data 1). The reliability, coverage and availability of a standalone version of all the software code were considered during the selection process. The reliability of the Rosetta protocol 3 employed in the Design module was benchmarked against experimental stability data previously collected for multiple-point mutants in our laboratory (46) as well as 1573 single-point mutants available in the ProTherm and HotMuSiC databases (Supplementary Data 2). These tests confirmed a significant correlation between half-lives and calculated changes in free energy $\Delta\Delta G$, as well as an ability of the fast protocol 3 to correctly



SELECT TYPE OF INPUT DATA

STRUCTURE

SEQUENCE

INPUT FROM SEQUENCE

Source : ● Enter own sequence ○ Enter Uniprot ID ○ Upload sequence file

Sequence :

```
MQDPYVKEAENLKKYFNAGHSDVADNGTLFLGILKN
WKEESDRKIMQSQIVSFYFKLFKNFKDDQSIQKSVETI
KEDMNV
KFFNSNKKKRDDFEKLTNYSVTDLNVQRKAIHELIQV
MAELSPAAKTGKRKRSQ
```

Search for structure

Structures :

	PDB ID	Resolution	
●	1FG9	2.9 Å	
○	1HIG	3.5 Å	
○	3BES	2.2 Å	

Process structure

INPUT FROM SEQUENCE

Source : ● Enter own sequence ○ Enter Uniprot ID ○ Upload sequence file

Sequence :

```
MKKILLLLVLIVAVLNFGKTIVTTINPYYLIVSQLLGDTAS
VKLLVPPGANPHLFSLKPSDAKTLEEADLIVANGLGLE
PYLEKYREKTVFVSDFIPALLLIDDNPHIWLDPFFLKYY
IVPGLYQVLIEKFPEKQSEIKQKAEEIVSGLDTVIRDSF
KALLPYTGKTVVMAHPSFTYFFKEFGLELITLSSGHEH
STSFSTIKEILRKKEQIVALFREPQQPAEILSSLEKELR
MKSFVLDPLGVNGEKTIVELLRKNLSVIQEALK
```

Search for structure

Modelling : ○ Download existing model ● Create new model

Modelling tool : ○ Modeller - faster & and less accurate (5 min) ○ I-Tasser - slower & more accurate (3 days)

Input : ● Use selected tool for template search and alignment ○ Enter own template ○ Enter own alignment

INPUT FROM SEQUENCE

Source : ● Enter own sequence ○ Enter Uniprot ID ○ Upload sequence file

Sequence :

```
    MKKILLLLVLIVAVLNFGKTIVTTINPYYLIVSQLLG
DTASVKLLVPPGANPHLFSLKPSDAKTLEEADLIVANG
LGLEPYLEKYREKTVFVSDFIPALLLIDDNPHIWLDPFF
LKYYIVPGLYQVLIEKFPEKQSEIKQKAEEIVSGLDTVI
RDSFKALLPYTGKTVVMAHPSFTYFFKEFGLELITLSS
GHEHSTSFSTIKEILRKKEQIVALFREPQQPAEILSSLE
KELRMKSFVLDPLGVNGEKTIVELLRKNLSVIQEALK
```

Search for structure

Modelling : ● Download existing model ○ Create new model

Models :

	model	provider	template	identity	from	to	coverage	reliability
●		MODBASE	1toaA	29 %	20	267	92 %	low
○		SWISSMODEL	2ps3.	27 %	20	267	92 %	low
○		NESG	2o1eA	20 %	20	266	92 %	low

Figure 2. Graphic user interface of the sequence input in the HotSpot Wizard 3.0. (**A**) Selection between structure and sequence input. (**B**) After entering of the sequence, searching for existing structures in PDB database is performed. (**C**) If no existing structure is found, search in homology model databases is performed. (**D**) Setting of homology modeling parameters—user can choose between Modeller and I-Tasser and eventually enter his own template or sequence alignment.

classify stabilizing and destabilizing mutations. Functionality of the Mutation design module was validated by saturation mutagenesis at the hotspot position L177 located at the tunnel mouth of the haloalkane dehalogenase LinB (47). Theoretical predictions correctly identified the variant L177W, which was found to be the most stable also experimentally (Supplementary Data 3). At last, we used the HotSpot Wizard 3.0 workflow for computational mutagenesis of six residues lining the active site cavity and the access tunnel of the haloalkane dehalogenases from non-pathogenic and pathogenic bacteria *Sphingobium japonicum* UT26 and *Mycobacterium tuberculosis* Rv2579, respectively (48). Single-point mutations and combined sixfold mutants were predicted using the automated protocols with crystal structures and homology models (Supplementary Data 4).

CONCLUSIONS AND OUTLOOK

HotSpot Wizard 3.0 is a new version of a popular web server used for the automated prediction of hotspots and the design of smart libraries in semi-rational protein design. In this version, homology modeling of the protein structure dramatically increases the usability of the platform by increasing the number of possible inputs and solves the limitation imposed by the number of available experimental structures. For homology modeling, Modeller and I-Tasser are used. The quality of the models created is evaluated using three different tools to identify wrongly modeled regions, which should be used for further computational design only with extreme care. The users are automatically warned whenever they attempt to redesign poorly resolved regions, for example the residues lying outside allowed regions of the Ramachandran plot. Rational design is further supported by the novel Mutation design module employing force field calculations for estimating the effect of substitution on protein thermodynamic stability. This new module can dramatically reduce the number of variants selected for experimental testing and can also help to pre-select mutations for identified positions during construction of smart libraries. In the future, we want to focus on more systematic use of multiple structural data from the Protein Data Bank, and on development of a novel engineering strategy for the design of biocatalysts that catalyze specific chemical reactions. Extensive databases searches will be coupled with the computational design module for identification of the best starting protein template for such an engineering exercise.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

Computational resources were supplied by the Ministry of Education, Youth and Sports of the Czech Republic under the Projects CESNET (Project No. LM2015042) and CERIT-Scientific Cloud (Project No. LM2015085).

FUNDING

Ministry of Education, Youth and Sports of the Czech Republic from the National Programme of Sustainability II [LQ1602, LQ1605, LO1214]; European Regional Development Fund [LM2015051, LM2015047, LM2015055]; Grant Agency of the Czech Republic [16–06096S]; European Union [720776, 722610]; Brno University Technology [FIT-S-17–3994 to L.S.]. Funding for open access charge: Czech Ministry of Education.

Conflict of interest statement. None declared.

REFERENCES

1. Hawkins,M.J., Soon-Shiong,P. and Desai,N. (2008) Protein nanoparticles as drug carriers in clinical medicine. *Adv. Drug Deliv. Rev.*, **60**, 876–885.
2. Godfrey,T. and Reichelt,J. (1982) Industrial applications. In: *Industrial Enzymology: The Application of Enzymes in Industry*. Macmillan, The Nature Press, London, pp. 582.
3. Bromley,E.H., Channon,K., Moutevelis,E. and Woolfson,D.N. (2008) Peptide and protein building blocks for synthetic biology: from programming biomolecules to self-organized biomolecular systems. *ACS Chem. Biol.*, **3**, 38–50.
4. De La Rica,R. and Matsui,H. (2010) Applications of peptide and protein-based materials in bionanotechnology. *Chem. Soc. Rev.*, **39**, 3499–3509.
5. Cheng,F., Zhu,L. and Schwaneberg,U. (2015) Directed evolution 2.0: improving and deciphering enzyme properties. *Chem. Commun.*, **51**, 9760–9772.
6. Romero,P.A. and Arnold,F.H. (2009) Exploring protein fitness landscapes by directed evolution. *Nat. Rev. Mol. Cell Biol.*, **10**, 866–876.
7. Lutz,S. (2010) Beyond directed evolution—semi-rational protein engineering and design. *Curr. Opin. Biotechnol.*, **21**, 734–743.
8. Cheng,Z., Peplowski,L., Cui,W., Xia,Y., Liu,Z., Zhang,J., Kobayashi,M. and Zhou,Z. (2017) Identification of key residues modulating the stereoselectivity of nitrile hydratase towards rac-Mandelonitrile by Semi-rational engineering. *Biotechnol. Bioeng.*, **115**, 1–12.
9. Bendl,J., Stourac,J., Sebestova,E., Vavra,O., Musil,M., Brezovsky,J. and Damborsky,J. (2016) HotSpot Wizard 2.0: automated design of site-specific mutations and smart libraries in protein engineering. *Nucleic Acids Res.*, **44**, W479–W487.
10. Talukdar,P. and Talapatra,S.N. (2017) Oxy-haemoglobin protein engineering: an automated design for hotspots stability, site-specific mutations and smart libraries by using HotSpot Wizard 2.0 software. *Int. J. Adv. Res. Comput. Sci.*, **8**, 220–228.
11. Wang,X., Ma,R., Xie,X., Liu,W., Tu,T., Zheng,F., You,S., Ge,J., Xie,H., Yao,B. *et al.* (2017) Thermostability improvement of a Talaromyces leycettanus xylanase by rational protein engineering. *Sci. Rep.*, **7**, 15287.
12. Vatansever,R., Uras,M.E., Sen,U., Ozyigit,I.I. and Filiz,E. (2016) Isolation of a transcription factor DREB1A gene from Phaseolus vulgaris and computational insights into its characterization: protein modeling, docking and mutagenesis. *J. Biomol. Struct. Dyn.* **35**, 1–12.
13. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.
14. UniProt Consortium. (2017) UniProt: the universal protein knowledgebase. *Nucleic Acids Res.*, **45**, D158–D169.
15. Baker,D. and Sali,A. (2001) Protein structure prediction and structural genomics. *Science*, **294**, 93–96.
16. Cavasotto,C.N. and Phatak,S.S. (2009) Homology modeling in drug discovery: current trends and applications. *Drug Discov. Today*, **14**, 676–683.
17. Schwede,T. (2013) Protein modeling: what happened to the ‘protein structure gap’? *Structure*, **21**, 1531–1540.
18. Haas,J., Roth,S., Arnold,K., Kiefer,F., Schmidt,T., Bordoli,L. and Schwede,T. (2013) The Protein Model Portal—a comprehensive resource for protein structure and model information. *Database*, **2013**, bat031.
19. Csmp.ucsf.edu. (2017) CSMP | Home. http://csmp.ucsf.edu/index.htm (20 December 2017, date last accessed).
20. Jcsg.org. (2017) The Joint Center for Structural Genomics (JCSG) Homepage. http://www.jcsg.org/ (20 December 2017, date last accessed).
21. Mcsg.anl.gov. (2017) http://www.mcsg.anl.gov/ (20 December 2017, date last accessed).
22. Nesg.org. (2017) NESG - NorthEast Structural Genomics consortium. http://www.nesg.org/ (20 December 2017, date last accessed).
23. Venkatagiriyappa,V. (2017) NYSGRC. http://www.nysgxrc.org/psi3-cgi/index.cgi (20 December 2017, date last accessed).
24. Jcmm.burnham.org. (2017) Joint Center for Molecular Modeling (JCMM). http://jcmm.burnham.org/ (20 December 2017, date last accessed).
25. Pieper,U., Webb,B.M., Dong,G.Q., Schneidman-Duhovny,D., Fan,H., Kim,S.J. and Tainer,J.A. (2013) ModBase, a database of annotated comparative protein structure models and associated resources. *Nucleic Acids Res.*, **42**, D336–D346.

26. Kiefer,F., Arnold,K., Künzli,M., Bordoli,L. and Schwede,T. (2008) The SWISS-MODEL repository and associated resources. *Nucleic Acids Res.*, **37**, D387–D392.
27. Biasini,M., Bienert,S., Waterhouse,A., Arnold,K., Studer,G., Schmidt,T. and Schwede,T. (2014) SWISS-MODEL: modelling protein tertiary and quaternary structure using evolutionary information. *Nucleic Acids Res.*, **42**, W252–W258.
28. Song,Y., DiMaio,F., Wang,R.Y.R., Kim,D., Miles,C., Brunette,T.J. and Baker,D. (2013) High-resolution comparative modeling with RosettaCM. *Structure*, **21**, 1735–1742.
29. Kim,D.E., Chivian,D. and Baker,D. (2004) Protein structure prediction and analysis using the Robetta server. *Nucleic Acids Res.*, **32**, W526–W531.
30. Kelley,L.A., Mezulis,S., Yates,C.M., Wass,M.N. and Sternberg,M.J. (2015) The Phyre2 web portal for protein modeling, prediction and analysis. *Nat. Protoc.*, **10**, 845–858.
31. Larsson,P., Skwark,M.J., Wallner,B. and Elofsson,A. (2010) Improved predictions by Pcons. net using multiple templates. *Bioinformatics*, **27**, 426–427.
32. Webb,B. and Sali,A. (2014) Protein structure modeling with MODELLER. *Methods Mol. Biol.*, **1137**, 151–115.
33. Yang,J., Yan,R., Roy,A., Xu,D., Poisson,J. and Zhang,Y. (2015) The I-TASSER Suite: protein structure and function prediction. *Nat. Methods*, **12**, 7–8.
34. McGuffin,L.J., Atkins,J.D., Salehe,B.R., Shuid,A.N. and Roche,D.B. (2015) IntFOLD: an integrated server for modelling protein structures and functions from amino acid sequences. *Nucleic Acids Res.*, **43**, W169–W173.
35. Russel,D., Lasker,K., Webb,B., Velázquez-Muriel,J., Tjioe,E., Schneidman-Duhovny,D. and Sali,A. (2012) Putting the pieces together: integrative modeling platform software for structure determination of macromolecular assemblies. *PLoS Biol.*, **10**, e1001244.
36. Hildebrand,A., Remmert,M., Biegert,A. and Söding,J. (2009) Fast and accurate automatic structure prediction with HHpred. *Proteins*, **77**, 128–132.
37. Källberg,M., Wang,H., Wang,S., Peng,J., Wang,Z., Lu,H. and Xu,J. (2012) Template-based protein structure modeling using the RaptorX web server. *Nat. Protoc.*, **7**, 1511–1522.
38. Yang,Y., Faraggi,E., Zhao,H. and Zhou,Y. (2011) Improving protein fold recognition and template-based modeling by employing probabilistic-based matching between predicted one-dimensional structural properties of query and corresponding native properties of templates. *Bioinformatics*, **27**, 2076–2082.
39. Kryshtafovych,A., Fidelis,K. and Moult,J. (2014) CASP10 results compared to those of previous CASP experiments. *Proteins*, **82**, 164–174.
40. Laskowski,R.A., MacArthur,M.W., Moss,D.S. and Thornton,J.M. (1993) PROCHECK: a program to check the stereochemical quality of protein structures. *J. Appl. Crystallogr.*, **26**, 283–291.
41. Chen,V.B., Arendall,W.B., Headd,J.J., Keedy,D.A., Immormino,R.M., Kapral,G.J. and Richardson,D.C. (2010) MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallogr. D Biol. Crystallogr.*, **66**, 12–21.
42. Hooft,R W., Vriend,G., Sander,C. and Abola,E.E. (1996) Errors in protein structures. *Nature*, **381**, 272–272.
43. Kellogg,E.H., Leaver-Fay,A. and Baker,D. (2011) Role of conformational sampling in computing mutation-induced changes in protein structure and stability. *Proteins*, **79**, 830–838.
44. Schymkowitz,J., Borg,J., Stricher,F., Nys,R., Rousseau,F. and Serrano,L. (2005) The FoldX web server: an online force field. *Nucleic Acids Res.*, **33**, W382–W388.
45. Kellogg,E.H., Leaver-Fay,A. and Baker,D. (2011) Role of conformational sampling in computing mutation-induced changes in protein structure and stability. *Proteins*, **79**, 830–838.
46. Bednar,D., Beerens,K., Sebestova,E., Bendl,J., Khare,S., Chaloupkova,R., Prokop,Z., Brezovsky,J., Baker,D. and Damborsky,J. (2015) FireProt: energy-and evolution-based computational design of thermostable multiple-point mutants. *PLoS Comput. Biol.*, **11**, e1004556.
47. Chaloupková,R., Sykorova,J., Prokop,Z., Jesenska,A., Monincova,M., Pavlova,M., Tsuda,M., Nagata,Y. and Damborsky,J. (2003) Modification of activity and specificity of haloalkane dehalogenase from Sphingomonas paucimobilis UT26 by engineering of its entrance tunnel. *J. Biol. Chem.*, **278**, 52622–52628.
48. Nagata,Y., Prokop,Z., Marvanova,S., Sykorova,J., Monincova,M., Tsuda,M. and Damborsky,J. (2003) Reconstruction of mycobacterial dehalogenase Rv2579 by cumulative mutagenesis of haloalkane dehalogenase LinB. *Appl. Environ. Microbiol.*, **69**, 2349–2355.

A.9 Paper IX

FireProt: web server for automated design of thermostable proteins

Published online 26 April 2017 *Nucleic Acids Research, 2017, Vol. 45, Web Server issue* ***W393–W399***
doi: 10.1093/nar/gkx285

FireProt: web server for automated design of thermostable proteins

Milos Musil[1,2,3,†], Jan Stourac[1,3,†], Jaroslav Bendl[1,2,3], Jan Brezovsky[1,3], Zbynek Prokop[1,3], Jaroslav Zendulka[2,4], Tomas Martinek[1,2,4], David Bednar[1,3,*] and Jiri Damborsky[1,3,*]

[1]Loschmidt Laboratories, Department of Experimental Biology, Masaryk University, Brno, Czech Republic, [2]Department of Information Systems, Faculty of Information Technology, Brno University of Technology, Brno, Czech Republic, [3]International Centre for Clinical Research, St. Anne's University Hospital Brno, Brno, Czech Republic and [4]Centre of Excellence IT4Innovations, Technical University Ostrava, Ostrava

Received February 11, 2017; Revised April 02, 2017; Editorial Decision April 10, 2017; Accepted April 11, 2017

ABSTRACT

There is a continuous interest in increasing proteins stability to enhance their usability in numerous biomedical and biotechnological applications. A number of *in silico* tools for the prediction of the effect of mutations on protein stability have been developed recently. However, only single-point mutations with a small effect on protein stability are typically predicted with the existing tools and have to be followed by laborious protein expression, purification, and characterization. Here, we present FireProt, a web server for the automated design of multiple-point thermostable mutant proteins that combines structural and evolutionary information in its calculation core. FireProt utilizes sixteen tools and three protein engineering strategies for making reliable protein designs. The server is complemented with interactive, easy-to-use interface that allows users to directly analyze and optionally modify designed thermostable mutants. FireProt is freely available at http://loschmidt.chemi.muni.cz/fireprot.

INTRODUCTION

Proteins are widely used in numerous biomedical and biotechnological applications. However, naturally occurring proteins cannot usually withstand the harsh industrial environment, since they are mostly evolved to function at mild conditions (1). Protein engineering has revolutionized the utilization of naturally available proteins for different industrial applications by improving various protein features such as stability, activity or enantioselectivity to surpass their natural limitations. Protein stability is generally strongly correlated with its expression yield (2), half-life (3), serum survival time (4) and performance in the presence of denaturing agents (5). Thus, stability is one of the key determinants of proteins applicability in biotechnological processes.

In the ideal case, the saturation mutagenesis would be applied to evaluate every possible mutation on every position of the engineered protein (6). However, such a search space would be enormous and the experimental evaluation can delay the design of truly thermostable protein for months or even years. Therefore, there are demands for effective and precise predictive computation of protein stability. To satisfy this goal a number of *in silico* tools have been developed recently. Some of these tools such as EASE-MM (7), I-Mutant (8) or mCSM (9) are based on machine learning techniques. Others are using so-called energetic functions. These programs can be further categorized into two groups. The first group utilizes a physical effective energy function for simulating the fundamental forces between atoms and is represented by the programs like Rosetta (10) and Eris (11). The second group is based on statistical potentials for which the energies are derived from frequencies of residues or atom contacts reported in the datasets of experimentally characterized protein mutants, e.g. PopMuSiC (12) and FoldX (13). However, due to the potentially antagonistic effect of mutations, only single-point mutations are usually predicted *in silico* and have to be followed by laborious and costly protein expression, purification and characterization. Single-point mutations typically enhance the melting temperature of target proteins by units of degree (3,14). A much higher degree of stabilization can be achieved by constructing multiple-point mutants (15). We have recently developed the FireProt (16), combining energy- and evolution-based approaches for reliable design of stable multiple-point mutants. The protocol includes several preceding filters that accelerate the calculation by omitting potentially deleterious mutations. FireProt is currently

[*]To whom correspondence should be addressed. Tel: +420 5 4949 3467; Fax: +420 5 4949 6302; Email: jiri@chemi.muni.cz; Website address: http://loschmidt.chemi.muni.cz/fireprot (type of web server: computational workflow)
Correspondence may also be addressed to David Bednar. Email: 222755@mail.muni.cz
[†]These authors contributed equally to the work as first authors.

available only in a stand-alone format and requires extensive experience in bioinformatics to carry out all necessary steps of the work flow. Currently, we are aware of only one server for design of stable multiple-point mutants - PROSS (17), utilizing Rosetta modeling and phylogenetic sequence information in its computation core.

Here, we present a web version of FireProt for the automated design of thermostable proteins. FireProt integrates sixteen computational tools and utilizes both sequence and structural information. FireProt web server provides users with thermostable proteins, constructed by three distinct strategies: (i) evolution-based approach, utilizing back-to-consensus analysis; (ii) energy-based approach, evaluating change in free energy upon mutation and (iii) combination of both evolution-based and energy-based approaches. In our view, it is very important to have this integrated approach, since phylogenetic analysis enables identification of the mutations stabilized by entropy, which cannot be predicted by force field calculations (Beerens *et al.*, under review). The server allows users to include preferred mutations into the thermostable protein, to generate corresponding structures and sequences for gene syntheses. Compared to the previously published FireProt protocol (16), minimum effort and no bioinformatics knowledge is required from users to calculate and analyze the results. Furthermore, all input parameters and computational protocols were optimized to minimize otherwise highly time demanding procedure. The server was complemented with a graphical interface allowing users to directly analyze the protein of interest and design multiple-point mutants.

MATERIALS AND METHODS

The basic workflow of FireProt strategy is outlined in Figure 1. In order to design a highly reliable thermostable multiple-point mutant, a protein defined by the user is annotated using several prediction tools and databases (Phase 1). With this knowledge in hand, energy- and evolution-based approach is applied to assemble a list of potentially stabilizing single-point mutations (Phase 2). Finally, three multiple-point mutants are generated in an additive manner, while removing potentially antagonistic effects of mutations (Phase 3).

Phase 1: Annotation of the protein

Initially, the user is requested to specify the protein structure, either by providing its PDB ID or by uploading a user-defined PDB file. The biological assembly of the target protein is then automatically generated by the MakeMultimer tool (http://watcut.uwaterloo.ca/tools/makemultimer/). Sequence homologs are obtained by performing a BLAST search (18) against the UniRef90 database (19), using the target protein sequence as an input query. Identified homologs are then aligned with the query protein using USEARCH (20), while sequences whose identity with the query is below or above the user defined thresholds (default: 30 and 90%) are excluded from the list of homologs. The remaining sequences are clustered using UCLUST (20), with a 90% identity threshold to remove close homologs. The cluster representatives are sorted based on the BLAST query coverage and by default, the first 200 of them are used to create a multiple sequence alignment with Clustal Omega tool (21). The multiple sequence alignment is used to: (i) estimate the conservation coefficient of each residue position in the protein based on the Jensen–Shannon entropy (22); (ii) identify correlated positions employing a consensual decision of the OMES (23), MI (24), aMIc (25), DCA (26), SCA (27), ELSC (28), McBASC (29) and (iii) analyze amino acid frequencies at individual positions within the protein.

Phase 2: Prediction of single-point mutations

In accordance with the original FireProt protocol, potentially stabilizing single-point mutations are identified via two separate branches: one relying on the estimation of the change of free energy upon mutation and second utilizing back-to-consensus approach.

The first, energy-based approach is employing FoldX and Rosetta tools that performed best on our testing dataset. Preceding filters accelerate the calculation by omitting potentially deleterious mutations. Prior to the identification of the single-point mutations itself, the target protein structure is amended and minimized. FoldX protocol is utilized to fill in the missing atoms in the residues and patched structure is consequently minimized with Rosetta minimization module. Conserved and correlated positions are immediately excluded from further analysis. It was observed that functional and structural constraints in proteins generally lead to the conservation of amino acid residues (30–33). Similarly, correlated residues ordinarily help to maintain protein function, folding or stability (34–36). Mutations conducted on these positions are therefore considered unsafe by current FireProt strategy, even though there is certainly a space for more sophisticated treatment of correlated positions, which will be further developed in future versions of FireProt server.

The remaining positions are subjected to saturation mutagenesis by using FoldX tool. Mutations with predicted ddG over given threshold (default: –1 kcal/mol) are steered away and rest is forwarded to Rosetta calculations. Finally, the mutations predicted by Rosetta as strongly stabilizing (default cut-off: –1 kcal/mol) are tagged as potential candidates for the design of the multiple-point mutants.

A high time demands of Rosetta analysis were one of the most excruciating issues with the original FireProt protocol. Even with the application of filters over 100 mutations was usually left for precise, but slow, Rosetta calculations. For this reason, we have evaluated several force fields and Rosetta protocols with the newly assembled dataset containing 1573 mutations from ProTherm database (37) and HotMuSiC dataset (38). Based on the results of the evaluations, the best trade-off between the time requirements and precision was selected. With Rosetta protocol 3, we have achieved more than tenfold increase in calculation speed while preserving high prediction accuracy. Details on dataset construction and protocols evaluation can be found in the Supplement 1 (Supplementary Tables S1–S5).

The second approach is based on the information obtained from multiple sequence alignment. The most common amino acid in each position of protein sequence often

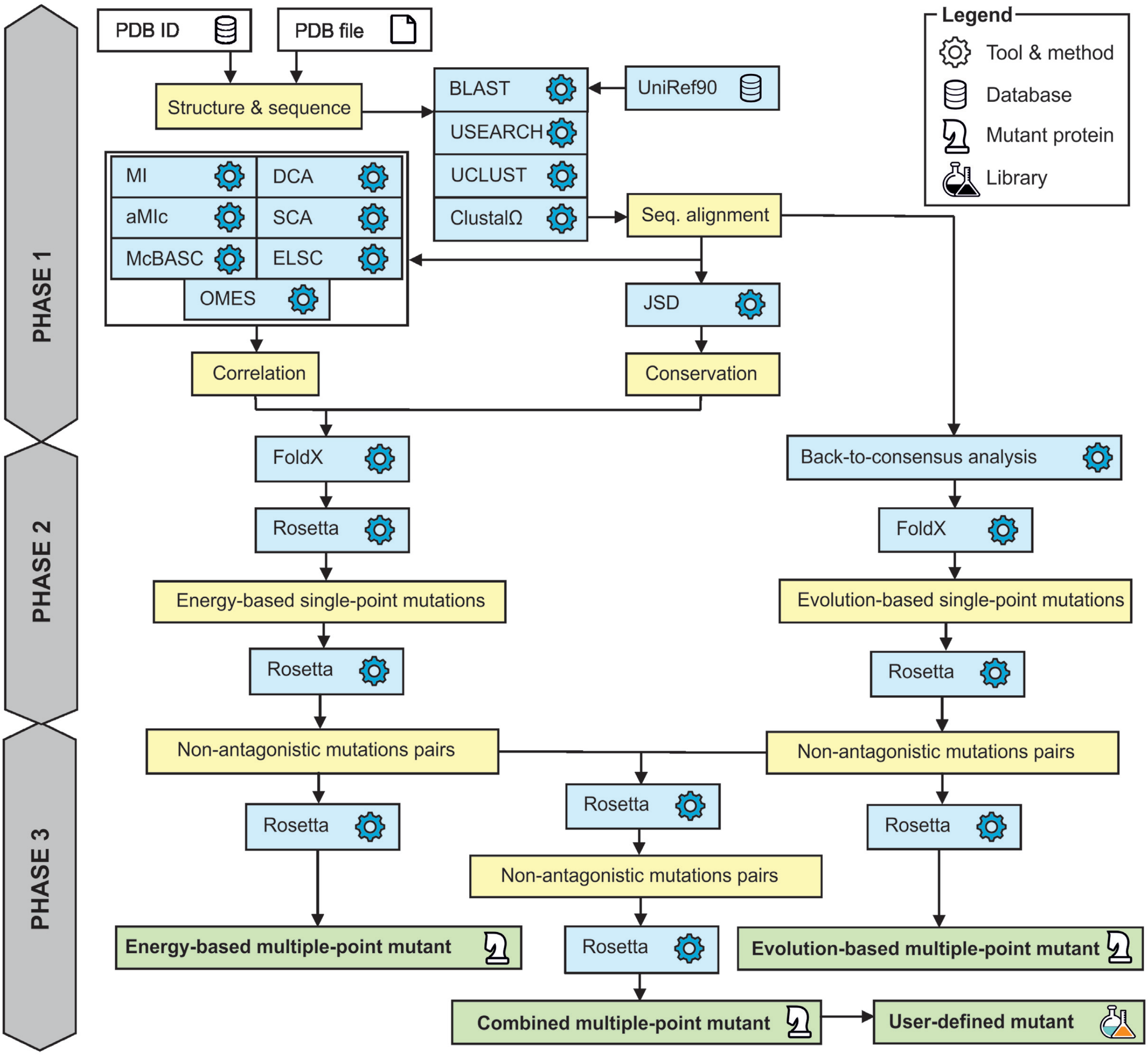

PHASE 1
PHASE 2
PHASE 3
PDB ID
PDB file
Structure & sequence
BLAST
USEARCH
UCLUST
ClustalΩ
UniRef90
Seq. alignment
MI
DCA
aMIc
SCA
McBASC
ELSC
OMES
JSD
Correlation
Conservation
FoldX
Rosetta
Energy-based single-point mutations
Rosetta
Non-antagonistic mutations pairs
Rosetta
Energy-based multiple-point mutant
Back-to-consensus analysis
FoldX
Evolution-based single-point mutations
Rosetta
Non-antagonistic mutations pairs
Rosetta
Evolution-based multiple-point mutant
Rosetta
Non-antagonistic mutations pairs
Rosetta
Combined multiple-point mutant
User-defined mutant
Legend
Tool & method
Database
Mutant protein
Library

Figure 1. Workflow of FireProt strategy.

provides a non-negligible effect on protein stability (39–42). Therefore, FireProt implements majority and frequency ratio approach to identify mutations at positions where the wild-type amino acid differs from the most prevalent one. By default, the single out mutations are located in the positions where the consensus residue is present in at least 50% of all analyzed sequences (majority method) or where consensus residue frequency is 40% and is at least five times more frequent than the wild-type amino acid (frequency ratio method). These thresholds were chosen in accordance to the previously published HotSpot Wizard method (43). Selected mutations are evaluated by FoldX and the stabilizing ones are listed as candidate mutations for the engineering of multiple-point mutant.

Phase 3: Design of thermostable protein

In total, three protein designs are provided by FireProt strategy. The first design includes only the mutations from energy-based approach, the second contains the mutations suggested by the evolution-based approach and the third is the combination of both. Naturally, because of potentially antagonistic effects between individual mutations, we cannot combine individual mutations blindly.

To avoid possible clashes, FireProt strategy is trying to minimize antagonistic effects by utilizing Rosetta. In the first step, all pairs of single-point mutations within the range of 10 Å are evaluated separately for energy- and evolution-based approach. Once change in free energy is obtained for all residue pairs, FireProt starts to introduce them into the multiple-point mutant in the order based on their predicted stability, excluding the mutations that are colliding with already included mutations. Algorithm stops once there are no mutations left or the stabilizing effect of analyzed pair drops below defined threshold.

Upon the completion of previous step, procedure is repeated this time considering only the pairs between the mutations chosen for the construction of energy- and evolution-based mutants. Finally, structures of all three mutants are modeled using the Rosetta protocol 16.

DESCRIPTION OF THE WEB SERVER

Input

The only required input to the web server is a tertiary structure of the protein of interest, provided either as a PDB ID or a user-defined PDB file. The user can then choose a predefined biological unit generated by the MakeMultimer tool or manually select chains for which the calculation should be performed. The calculations can be configured in either basic or advanced mode.

In the basic mode, user is allowed to change the setting of BLAST search and alignment construction. The advanced mode expands the list of modifiable parameters by the ones connected with: (i) the identification of consensus residues by majority and frequency ratio approach, (ii) the thresholds used by FoldX and Rosetta prediction tools and (iii) the decision threshold employed in the consensual analysis of correlated positions. Advanced mode allows expert users to fine-tune the parameters of calculation according to studied systems. However, the presented default values are optimized to provide reliable results for most of the systems and we therefore do not advice their change in the general scenarios.

Output

Upon submission, a unique identifier is assigned to each job to track the calculation and the 'Results browser' informs the user about the status of the individual steps in the FireProt workflow (Figure 2B). Once the job is finished, users can either directly download the results in the .zip archive or navigate themselves into the 'Results page' for further analysis. The 'Results page' is intuitively organized into several panels as described below.

Protein visualization. The wild-type and the mutant structure is interactively visualized in the web browser (Figure 2D) utilizing the Jsmol applet (http://wiki.jmol.org/index.php/JSmol). Users can switch between different protein visualization styles and also highlight selected amino acids in the protein structure. Residues that were included into energy-based mutant are colored in orange, evolution-based mutations are in blue and all other residues are in gray. User selected residues that were not part of any mutant are underlined in red.

Mutant overview. The 'Mutant overview' panel is organized into four tabs (Figure 2A). The first three tabs provide information about mutations included into combined, energy-based and evolution-based mutant. The checkbox, allowing users to visualize the chosen residues in Jsmol applet, can be found in each row together with all data relevant for a given computational approach. The last tab contains the list of all residues in the wild-type structure. While 'wild-type' tab is active, the wild-type structure is visualized in Jsmol applet instead of the mutated one and the user is allowed to introduce user-defined mutations into multiple-point mutant via the 'plus' icon in the last column.

General information. The 'FireProt protocol design' panel provides users with general information about the target protein and the designs constructed by FireProt strategy, such as a number of mutations and estimated change in free energy (Figure 2C).

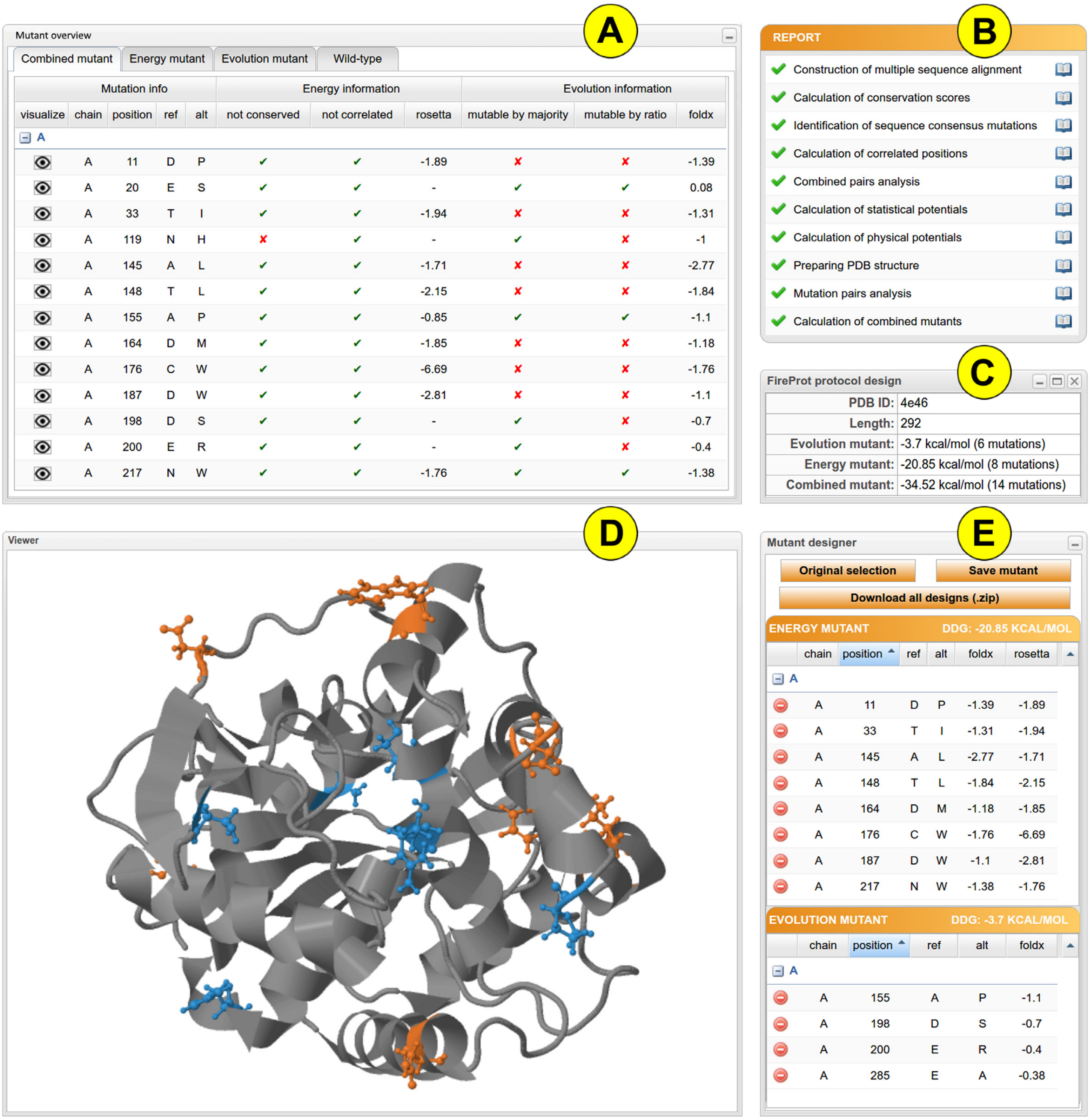
Mutant designer. The 'Mutant designer' panel allows the user to design own multiple-point mutant by managing mutations divided into energy- and evolution-based subset. If all mutations in the subset have their predicted energy values assigned, a total change in Gibbs free energy is immediately estimated assuming simple additivity. Users can also generate an amino acid sequence from the designed multiple-point mutant that combines mutations included into energy- and evolution-based subsets. All prepared designs can be downloaded in one .zip archive (Figure 2E).

EXPERIMENTAL VALIDATION

The original FireProt strategy was experimentally verified with three proteins (haloalkane dehalogenase DhaA, PDB ID 4E46; γ-hexachlorocyclohexane dehydrochlorinase LinA, PDB ID 3A76; and fibroblast growth factor 2, PDB ID 4OEE) and provided respective stabilization of proteins $\Delta T_\mathrm{m} = 25$, 21 and 15°C (Table 1). The original protocol was modified to enable fully automated calculation at the reasonable time, while maintaining high prediction accuracy (Supplementary Table S6). Prediction of eight multiple-point mutants using this modified protocol was validated using the data of FRESCO (44) and identified mutations were compared with another online protein stabilization tool PROSS (17). FireProt and PROSS showed similar predictive power, correctly identifying 29 and 20 potentially stabilizing positions, respectively (Supplementary Table S7).

CONCLUSIONS AND OUTLOOK

FireProt is a web server that provides users with a one-stop-shop solution for the design of thermostable multiple-point mutant proteins. In comparison with the standalone FireProt strategy (16), all default parameters and computational protocols were optimized to increase the calculation speed, while maintaining the prediction accuracy. The designs produced by the FireProt workflow were experimentally verified and thus users can obtain highly reliable thermostable proteins with minimal experimental effort. The server is complemented by an easy-to-use graphical interface that allows users to interactively analyze individual mutations selected as a part of energy- or evolution-based approach together with the ability to design their own multiple-point mutants on top of our robust strategy.



Mutation info					Energy information			Evolution information		
visualize	chain	position	ref	alt	not conserved	not correlated	rosetta	mutable by majority	mutable by ratio	foldx
⊟ A										
👁	A	11	D	P	✓	✓	-1.89	✗	✗	-1.39
👁	A	20	E	S	✓	✓	-	✓	✓	0.08
👁	A	33	T	I	✓	✓	-1.94	✗	✗	-1.31
👁	A	119	N	H	✗	✓	-	✓	✗	-1
👁	A	145	A	L	✓	✓	-1.71	✗	✗	-2.77
👁	A	148	T	L	✓	✓	-2.15	✗	✗	-1.84
👁	A	155	A	P	✓	✓	-0.85	✓	✓	-1.1
👁	A	164	D	M	✓	✓	-1.85	✗	✗	-1.18
👁	A	176	C	W	✓	✓	-6.69	✗	✗	-1.76
👁	A	187	D	W	✓	✓	-2.81	✗	✗	-1.1
👁	A	198	D	S	✓	✓	-	✓	✗	-0.7
👁	A	200	E	R	✓	✓	-	✓	✗	-0.4
👁	A	217	N	W	✓	✓	-1.76	✓	✓	-1.38

- ✓ Construction of multiple sequence alignment
- ✓ Calculation of conservation scores
- ✓ Identification of sequence consensus mutations
- ✓ Calculation of correlated positions
- ✓ Combined pairs analysis
- ✓ Calculation of statistical potentials
- ✓ Calculation of physical potentials
- ✓ Preparing PDB structure
- ✓ Mutation pairs analysis
- ✓ Calculation of combined mutants

FireProt protocol design

PDB ID:	4e46
Length:	292
Evolution mutant:	-3.7 kcal/mol (6 mutations)
Energy mutant:	-20.85 kcal/mol (8 mutations)
Combined mutant:	-34.52 kcal/mol (14 mutations)

ENERGY MUTANT DDG: -20.85 KCAL/MOL

	chain	position	ref	alt	foldx	rosetta
⊟ A						
⊖	A	11	D	P	-1.39	-1.89
⊖	A	33	T	I	-1.31	-1.94
⊖	A	145	A	L	-2.77	-1.71
⊖	A	148	T	L	-1.84	-2.15
⊖	A	164	D	M	-1.18	-1.85
⊖	A	176	C	W	-1.76	-6.69
⊖	A	187	D	W	-1.1	-2.81
⊖	A	217	N	W	-1.38	-1.76

EVOLUTION MUTANT DDG: -3.7 KCAL/MOL

	chain	position	ref	alt	foldx
⊟ A					
⊖	A	155	A	P	-1.1
⊖	A	198	D	S	-0.7
⊖	A	200	E	R	-0.4
⊖	A	285	E	A	-0.38

Figure 2. FireProt's graphical user interface showing the results obtained for the haloalkane dehalogenase DhaA (PDB ID: 4e46). (**A**) The 'Mutant overview' panel provides a list of mutations introduced into protein structure. (**B**) The 'Report' panel shows the status of calculation in the individual steps of the computational pipeline. (**C**) The 'Protocol design' panel provides general information about FireProt designs. (**D**) The JSmol ´Viewer´ allows interactive visualization of the protein. (**E**) The 'Mutant designer' panel enables manual adjustment of a new combined mutant.

Table 1. Experimental validation of FireProt strategy

Protein PDB ID	Energy-based mutations	Evolution-based mutations	ΔT_m [°C]
4E46	8	3	+25
3A76	4	3	+21
4OEE	4	2	+15

The automation of the whole procedure makes the process of the design of thermostable proteins accessible to users without any prior expertise in bioinformatics since it eliminates the need to select, install and evaluate tools, optimize their parameters, and interpret intermediate results. However, the energy-based approach of the FireProt strategy depends on the quality of provided protein structure and therefore the prediction accuracy might be compromised in the case of low-resolution structures or homology models.

In the future, we plan to implement new strategies such as a design based on the analysis of correlated positions that would contribute to the construction of the final combined mutant, elimination of highly flexible regions and introduction of disulfide bridges. Also, we plan to equip FireProt with several new filters, e.g. exclusion of the amino acids located in the close neighborhoods of the active sites or the ones participating in oligomerization.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We would like to thank the developers of PROSS Adi Goldenzweig and Dr Sarel Fleishman (Weizmann Institute of Science, Israel) for providing the data set for validation purposes. Computational resources were supplied by the Ministry of Education, Youth and Sports of the Czech Republic under the Projects CESNET (Project No. LM2015042) and CERIT-Scientific Cloud (Project No. LM2015085) provided within the program Projects of Large Research, Development and Innovations Infrastructures.

FUNDING

This research and open access charge was funded by Ministry of Education, Youth and Sports of the Czech Republic from the National Programme of Sustainability II [LQ1605, LO1214]; European Regional Development Fund [ELIXIR-CZ LM2015047]; Grant Academy of the Czech Republic [16-06096S]; Brno University Technology [FIT-S-17-3964].

Conflict of interest statement. None declared.

REFERENCES

1. Modarres,H.P., Mofrad,M.R. and Sanati-Nezhad,A. (2016) Protein thermostability engineering. *RSC Adv.*, **6**, 115252–115270.
2. Ferdjani,S., Ionita,M., Roy,B., Dion,M., Djeghaba,Z., Rabiller,C. and Tellier,C. (2011) Correlation between thermostability and stability of glycosidases in ionic liquid. *Biotechnol. Lett.*, **33**, 1215–1219.
3. Wijma,H.J., Floor,R.J. and Janssen,D.B. (2013) Structure- and sequence-analysis inspired engineering of proteins for enhanced thermostability. *Curr. Opin. Struct. Biol.*, **23**, 588–594.
4. Gao,D., Narasimhan,D.L., Macdonald,J., Brim,R., Ko,M.C., Landry,D.W., Woods,J.H., Sunahara,R.K. and Zhan,C.G. (2009) Thermostable variants of cocaine esterase for long-time protection against cocaine toxicity. *Mol. Pharmacol.*, **75**, 318–323.
5. Polizzi,K.M., Bommarius,A.S., Broering,J.M. and Chaparro-Riggers,J.F. (2007) Stability of biocatalysts. *Curr. Opin. Chem. Biol.*, **11**, 220–225.
6. Gray,K.A., Richardson,T.H., Kretz,K., Short,J.M., Bartnek,F., Knowles,R., Kan,L., Swanson,P.E. and Robertson,D.E. (2001) Rapid evolution of reversible denaturation and elevated melting temperature in a microbial haloalkane dehalogenase. *Adv. Synth. Catal.*, **343**, 607–617.
7. Folkman,L., Stantic,B., Sattar,A. and Zhou,Y. (2016) EASE-MM: sequence-based prediction of mutation-induced stability changes with feature-based multiple models. *J. Mol. Biol.*, **428**, 1394–1405.
8. Capriotti,E., Fariselli,P. and Casadio,R. (2005) I-Mutant2.0: predicting stability changes upon mutation from the protein sequence or structure. *Nucleic Acids Res.*, W306–W310.
9. Pires,D.E., Ascher,D.B. and Blundell,T.L. (2014) mCSM: predicting the effects of mutations in proteins using graph-based signatures. *Bioinformatics*, **30**, 335–342.
10. Kellogg,E.H., Leaver-Fay,A. and Baker,D. (2011) Role of conformational sampling in computing mutation-induced changes in protein structure and stability. *Proteins*, **79**, 830–838.
11. Yin,S., Ding,F. and Dokholyan,N.V. (2007) Modeling backbone flexibility improves protein stability estimation. *Structure*, **15**, 1567–1576.
12. Dehouck,Y., Grosfils,A., Folch,B., Gilis,D., Bogaerts,P. and Rooman,M. (2009) Fast and accurate predictions of protein stability changes upon mutations using statistical potentials and neural networks: PoPMuSiC-2.0. *Bioinformatics*, **25**, 2537–2543.
13. Guerois,R., Nielsen,J.E. and Serrano,L. (2002) Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. *J. Mol. Biol.*, **320**, 369–387.
14. Gumulya,Y. and Reetz,M.T. (2011) Enhancing the thermal robustness of an enzyme by directed evolution: least favorable starting points and inferior mutants can map superior evolutionary pathways. *ChemBioChem*, **12**, 2502–2510.
15. Bommarius,A.S. and Paye,M.F. (2013) Stabilizing biocatalysts. *Chem. Soc. Rev.*, **42**, 6534–6565.
16. Bednar,D., Beerens,K., Sebestova,E., Bendl,J., Khare,S., Chaloupkova,R., Prokop,Z., Brezovsky,J., Baker,D. and Damborsky,J. (2015) FireProt: energy- and evolution-based computational design of thermostable multiple-point mutants. *PLoS Comput. Biol.*, **11**, e1004556.
17. Goldenzweig,A., Goldsmith,M., Hill,S.E., Gertman,O., Laurino,P., Ashani,Y., Dym,O., Unger,T., Albeck,S., Prilusky,J. *et al.* (2016) Automated structure- and sequence-based design of proteins for high bacterial expression and stability. *Mol. Cell*, **63**, 337–346.
18. Camacho,C., Coulouris,G., Avagyan,V., Ma,N., Papadopoulos,J., Bealer,K. and Madden,T.L. (2009) BLAST+: architecture and applications. *BMC Bioinformatics*, **10**, 421.
19. Suzek,B.E., Wang,Y., Huang,H., McGarvey,P.B. and Wu,C.H. (2015) UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics*, **31**, 926–932.
20. Edgar,R.C. (2010) Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*, **26**, 2460–2461.
21. Sievers,F., Wilm,A., Dineen,D., Gibson,T.J., Karplus,K., Li,W., Lopez,R., McWilliam,H., Remmert,M., Soding,J. *et al.* (2011) Fast, scalable generation of high-quality protein multiple sequence alignments using clustal omega. *Mol. Syst. Biol.*, **7**, 539.
22. Capra,J.A. and Singh,M. (2007) Predicting functionally important residues from sequence conservation. *Bioinformatics*, **23**, 1875–1882.
23. Kass,I. and Horovitz,A. (2002) Mapping pathways of allosteric communication in GroEL by analysis of correlated mutations. *Proteins*, **48**, 611–617.
24. Korber,B.T.M., Farber,R.M., Wolpert,D.H. and Lapedes,A.S. (1993) Covariation of mutations in the V3 loop of human immunodeficiency virus type 1 envelope protein: an information theoretic analysis. *Proc. Natl. Acad. Sci. U.S.A.*, **90**, 7176–7180.
25. Lee,B.C. and Kim,D. (2009) A new method for revealing correlated mutations under the structural and functional constraints in proteins. *Bioinformatics*, **25**, 2506–2513.
26. Weigt,M., White,R.A., Szurmant,H., Hoch,J.A. and Hwa,T. (2008) Identification of direct residue contacts in protein–protein interaction by message passing. *Proc. Natl. Acad. Sci. U.S.A.*, **106**, 67–72.
27. Lockless,S.W. and Ranganathan,R. (1999) Evolutionarily conserved pathways of energetic connectivity in protein families. *Science*, **286**, 295–299.
28. Dekker,J.P., Fodor,A., Aldrich,R.W. and Yellen,G. (2004) A perturbation-based method for calculating explicit likelihood of

evolutionary co-variance in multiple sequence alignments. *Bioinformatics*, **20**, 1565–1572.

29. Valencia,A. (2003) Multiple sequence alignments as tools for protein structure and function prediction. *Compar. Funct. Genomics*, **4**, 424–427.
30. Benner,S.A. and Gerloff,D. (1991) Patterns of divergence in homologous proteins as indicators of secondary and tertiary structure: a prediction of the structure of the catalytic domain of protein kinases. *Adv. Enzyme Regul.*, **31**, 121–181.
31. Brenner,S. (1988) The molecular evolution of genes and proteins: a tale of two serines. *Nature*, **334**, 528–530.
32. Cooperman,B.S., Baykov,A.A. and Lahti,R. (1992) Evolutionary conservation of the active site of soluble inorganic pyrophosphatase. *Trends Biochem. Sci.*, **17**, 262–266.
33. Howell,N. (1989) Evolutionary conservation of protein regions in the protonmotive cytochrome b and their possible roles in redox catalysis. *J. Mol. Evol.*, **29**, 157–169.
34. Gobel,U., Sander,C., Schneider,R. and Valencia,A. (1994) Correlated mutations and residue contacts in proteins. *Proteins*, **18**, 309–317.
35. Neher,E. (1994) How frequent are correlated changes in families of protein sequences? *Proc. Natl. Acad. Sci. U.S.A.*, **91**, 98–102.
36. Taylor,W.R. and Hatrick,K. (1994) Compensating changes in protein multiple sequence alignments. *Protein Eng.*, **7**, 341–348.
37. Bava,K.A., Gromiha,M.M., Uedaira,H., Kitajima,K. and Sarai,A. (2004) ProTherm, version 4.0: thermodynamic database for proteins and mutants. *Nucleic Acids Res.*, **32**, D120–D121.
38. Pucci,F., Bourgeas,R. and Rooman,M. (2016) Predicting protein thermal stability changes upon point mutations using statistical potentials: introducing HoTMuSiC. *Scientific Rep.*, **6**, 23257.
39. Amin,N., Liu,A.D., Ramer,S., Aehle,W., Meijer,D., Metin,M., Wong,S., Gualfetti,P. and Schellenberger,V. (2004) Construction of stabilized proteins by combinatorial consensus mutagenesis. *Protein Eng. Des. Select.*, **17**, 787–793.
40. Lehmann,M., Loch,C., Middendorf,A., Studer,D., Lassen,S.F., Pasamontes,L., van Loon,A.P. and Wyss,M. (2002) The consensus concept for thermostability engineering of proteins: further proof of concept. *Protein Eng.*, **15**, 403–411.
41. Pey,A.L., Rodriguez-Larrea,D., Bomke,S., Dammers,S., Godoy-Ruiz,R., Garcia-Mira,M.M. and Sanchez-Ruiz,J.M. (2008) Engineering proteins with tunable thermodynamic and kinetic stabilities. *Proteins*, **71**, 165–174.
42. Sullivan,B.J., Nguyen,T., Durani,V., Mathur,D., Rojas,S., Thomas,M., Syu,T. and Magliery,T.J. (2012) Stabilizing proteins from sequence statistics: the interplay of conservation and correlation in triosephosphate isomerase stability. *J. Mol. Biol.*, **420**, 384–399.
43. Bendl,J., Stourac,J., Sebestova,E., Vavra,O., Musil,M., Brezovsky,J. and Damborsky,J. (2016) HotSpot wizard 2.0: automated design of site-specific mutations and smart libraries in protein engineering. *Nucleic Acids Res.*, **44**, W479–W487.
44. Floor,R.J.1, Wijma,H.J., Colpa,D.I., Ramos-Silva,A., Jekel,P.A., Szymański,W., Feringa,B.L., Marrink,S.J. and Janssen,D.B. (2014) Computational library design for increasing haloalkane dehalogenase stability. *ChemBioChem*, **15**, 1660–1672.

A.10 Paper X

digIS: towards detecting distant and putative novel insertion sequence elements in prokaryotic genomes

BMC Bioinformatics

SOFTWARE Open Access

digIS: towards detecting distant and putative novel insertion sequence elements in prokaryotic genomes

Janka Puterová and Tomáš Martínek*

*Correspondence:
martinto@fit.vutbr.cz
IT4Innovations Centre
of Excellence, Faculty
of Information Technology,
Brno University
of Technology, Bozetechova
2, 612 66 Brno, Czechia

Abstract

Background: The insertion sequence elements (IS elements) represent the smallest and the most abundant mobile elements in prokaryotic genomes. It has been shown that they play a significant role in genome organization and evolution. To better understand their function in the host genome, it is desirable to have an effective detection and annotation tool. This need becomes even more crucial when considering rapid-growing genomic and metagenomic data. The existing tools for IS elements detection and annotation are usually based on comparing sequence similarity with a database of known IS families. Thus, they have limited ability to discover distant and putative novel IS elements.

Results: In this paper, we present *digIS*, a software tool based on profile hidden Markov models assembled from catalytic domains of transposases. It shows a very good performance in detecting known IS elements when tested on datasets with manually curated annotation. The main contribution of *digIS* is in its ability to detect distant and putative novel IS elements while maintaining a moderate level of false positives. In this category it outperforms existing tools, especially when tested on large datasets of archaeal and bacterial genomes.

Conclusion: We provide *digIS*, a software tool using a novel approach based on manually curated profile hidden Markov models, which is able to detect distant and putative novel IS elements. Although *digIS* can find known IS elements as well, we expect it to be used primarily by scientists interested in finding novel IS elements. The tool is available at https://github.com/janka2012/digIS.

Keywords: IS elements, Mobile element, Profile HMM, Prokaryotic genomes, Genome annotation

Background

Insertion sequence elements (IS elements) are the smallest and most abundant autonomous transposable elements in prokaryotic genomes, usually ranging from 700 bp to 3 kbp. However, there are exceptions, and some IS families (Tn3) can contain elements having a length greater than 5 kbp. ISs are widespread in prokaryotic genomes and may occur in high copy numbers. They play an essential role in genome evolution, structure, and host-genome adaptability. Due to their movement ability, IS elements represent mutagenic agents and can: cause modulation of expression of neighboring genes, affect virulence, change xenobiotic or antimicrobial resistance, or modulate metabolic activities. Detailed information on IS element function in host genomes can be found in recent reviews [1, 2].

Typically, IS elements consist of one or two open reading frames (ORFs) encoding a transposase (Tpase), a product necessary for transposition within a particular genome or horizontally between genomes (in plasmids). They are flanked by short terminal inverted repeats (IRs) and direct repeats (DRs). Transposases occurring in IS elements include five groups named after amino acid residues located at their conserved catalytic domain that catalyzes the transposition: DDE, DEDD, HUH, Tyrosine (Y), and Serine (S). IS elements with DDE transposase are the most abundant, and their conserved catalytic domain has a typical secondary structure $\beta 1 - \beta 2 - \beta 3 - \alpha 1 - \beta 4 - \alpha 2/3 - \beta 5 - \alpha 4 - \alpha 5/6$. Classification of IS elements into families is based mainly on Tpase structure, but other features such as IRs and DRs are also considered. Up to now, 29 IS families have been identified [1].

ISfinder [3] is a human-curated database and the most comprehensive source of known IS elements at present. Currently, the database contains more than 5000 entries and is updated regularly. As an extension of the ISfinder database, the authors implemented an ISbrowser interface [4] for visualization of IS elements inside genomes, and they prepared a benchmark dataset, consisting of 118 manually annotated prokaryotic genomes (as of November 2017), that is often used for assessment of IS detection tools performance. Another data source focused on mobile genetic elements, including manually annotated insertion sequences, is ACLAME database [5]. Unfortunately, this database has not been updated since 2009.

Even though the databases of known IS elements are growing, we are probably far from having a complete knowledge of all IS families and their structures. Therefore, for a better understanding of the IS elements function and their role in genome evolution, it is desirable to have an effective tool capable of not only annotating known families but also detecting new ones. This need becomes even more crucial when considering rapid-growing genomic and metagenomic data.

At present, there are several tools available for the detection of IS elements in prokaryotic genomes. Some of them are designed for searching in raw sequenced data (ISQuest [6], ISMapper [7], ISseeker [8], panISa [9]), and the others require assembled sequences (IScan [10], ISsaga [11], OASIS [12], ISEScan [13], TnpPred [14]). Almost all tools utilize a homology-based approach and are dependent on a source of known IS elements (they use a reference database either for verifying their results or for building searching profiles). Only the panISa tool detects IS elements solely based on structural features, such as an alignment of DR regions, and does not require a reference database.

Homology-based methods can be further divided into two main categories: (1) sequence-based and (2) profile-based methods. The first category is represented by tools IScan, OASIS, ISQuest, and ISseeker, which utilize the ISfinder database as a reference library in combination with BLAST software [15] to find close homologs. These tools are often used in annotation pipelines, where outputs with a high level of confidence are required.

The latter category includes ISsaga, TnpPred, and ISEScan. They take advantage of interpolated Markov models or profile hidden Markov models (pHMMs), which provide a more sensitive search, and detect remote homology sequences. ISsaga utilizes GLIMMER [16] and detects ORFs of IS elements or their fragments using an optimized interpolated Markov model built from the ISfinder database. TnpPred is focused on transposases detection (not full-length IS elements) and provides pHMMs for 19 of 29 IS families only. ISEScan uses 621 pHMMs built automatically from Tpases in the ACLAME database, but 355 of them are made up of one sequence only. Based on the configuration, ISEScan searches for whole Tpases or allow the presence of fragments.

Both sequence-based and profile-based tools can find new members of existing IS families, as they usually share significant sequence similarity either at the DNA or Tpase/ORF level. Profile-based methods are able to find remote members with lower similarity, which can represent hitherto undiscovered families—distant putative novel IS elements. However, the reliable identification of new IS families and their members is still challenging even for existing profile-based tools. It is mainly due to the Tpase structure, which comprises of several, often variable, domains. A search for the whole Tpase (ISEScan) is quite specific and unable to uncover novel IS elements with a distinct Tpase structure. On the other hand, allowing for fragments (ISEScan, ISsaga, and TnpPred) may result in many hits having significant similarity to a specific part of a completely different protein (i.e., false positives in terms of tool evaluation).

In this paper, we address the aforementioned challenge using a novel approach to detecting distant members of known IS families and putative novel IS elements. The fundamental idea is to search for the most conserved part of Tpase—the catalytic domain. The search is based on manually curated pHMMs with noise cutoff thresholds. Utilizing this approach, we can detect both known and putative novel IS elements with a moderate level of false positives while maintaining high sensitivity. The proposed method is implemented as *digIS* software and released as open-source at https://github.com/janka2012/digIS. The installed tool, including all dependencies, is also available as a docker image at https://hub.docker.com/r/janka2012/digis.

Implementation

digIS is a command-line tool developed in Python. It utilizes several external tools such as BLAST [15], HMMER [17], and Biopython library [18]. As an input, *digIS* accepts contigs in FASTA format. Optionally, the user can provide a GenBank annotation file for a given input sequence(s). This annotation is later used to improve the classification of identified IS elements (see "Output classification" section).

Firstly, we built a library of manually curated pHMMs, corresponding to Tpase catalytic domains of individual IS families. As a source of sequences, we used the ISfinder database, and for each pHMM, we identified the noise cutoff threshold.

Then, the *digIS* search pipeline operates in the following way:

1 The whole input nucleic acid sequence is translated into amino acid sequences (all six frames).
2 The translated sequences are searched using manually curated pHMMs.

3 Found hits, referred to as *seeds*, are filtered by domain bit score and e-value. Those that overlap or follow one another within a certain distance are merged.
4 Each seed is matched against the database of known IS elements (ISfinder) and its genomic positions are extended according to the best hit.
5 Extended seeds are filtered by noise cutoff score and length. Duplicates, corresponding to the same IS element, are removed.
6 Remaining extended seeds are classified based on sequence similarity and GenBank annotation (if available) to assess their quality.
7 Finally, the classified outputs are reported in the CSV and GFF3 format.

The overall *digIS* workflow is depicted in Fig. 1, and the individual steps are described in detail in the following sections.

Building profile hidden Markov models for the transposase catalytic domain of individual IS families

Tpase sequences were obtained from the ISfinder database. For each IS family, the pHMM was created as follows: (1) the longest ORF sequence, representing Tpase and its catalytic domain, was chosen for each IS element[1], (2) a multiple sequence alignment (MSA) for a set of Tpases belonging to the same family was created by Clustal Omega [19] and visualized using Jalview [20], (3) for each MSA, a protein secondary structure of the transposase was predicted using JPred4 [21] and used to determine the boundaries of the conserved catalytic core; the MSA was refined based on the positions of the catalytic residues (usually DDE), and the catalytic domain was manually cut using these determined boundaries, (4) such a manually modified MSA was used to construct resultant pHMM using *hmmbuild* from the HMMER package.

Since IS3, IS4, and IS5 families contain multiple subfamilies, a separate model was constructed for each of them. Moreover, IS5/IS5 and IS5/None subfamilies showed various sequence patterns (e.g., long insertions, deletions), and therefore several models were built for them concerning these patterns. MSAs with highlighted sequence groups used to construct these models are available in Additional files 1 and 2. For the ISNCY family, models were built for IS1202 and ISDol1 subfamilies only, since other subfamilies did not contain a sufficient amount of sequences. We required the models to be assembled from at least ten sequences to have a generalizing ability to find distant Tpases. Altogether, 50 pHMMs were constructed.

The remaining sequences of IS5 and ISNCY subfamilies representing outliers/distant sequences were cut with regard to the catalytic residues and secondary structure. They were used later as *individual* protein sequences in *phmmer* search. Overall, 70 outlier sequences were collected.

To eliminate false-positive hits reported by HMMER using pHMMs and still have the ability to detect distant and novel IS elements, a domain noise cutoff threshold—which represents a bit score of the highest-scoring known false positive—was

[1] Various IS families carry Tpase consisting of multiple ORFs. These ORFs are present in the ISfinder database in both individual and fusion forms. As duplicated sequences may lead to a bias in pHMMs, only the longest ORF sequence was used.

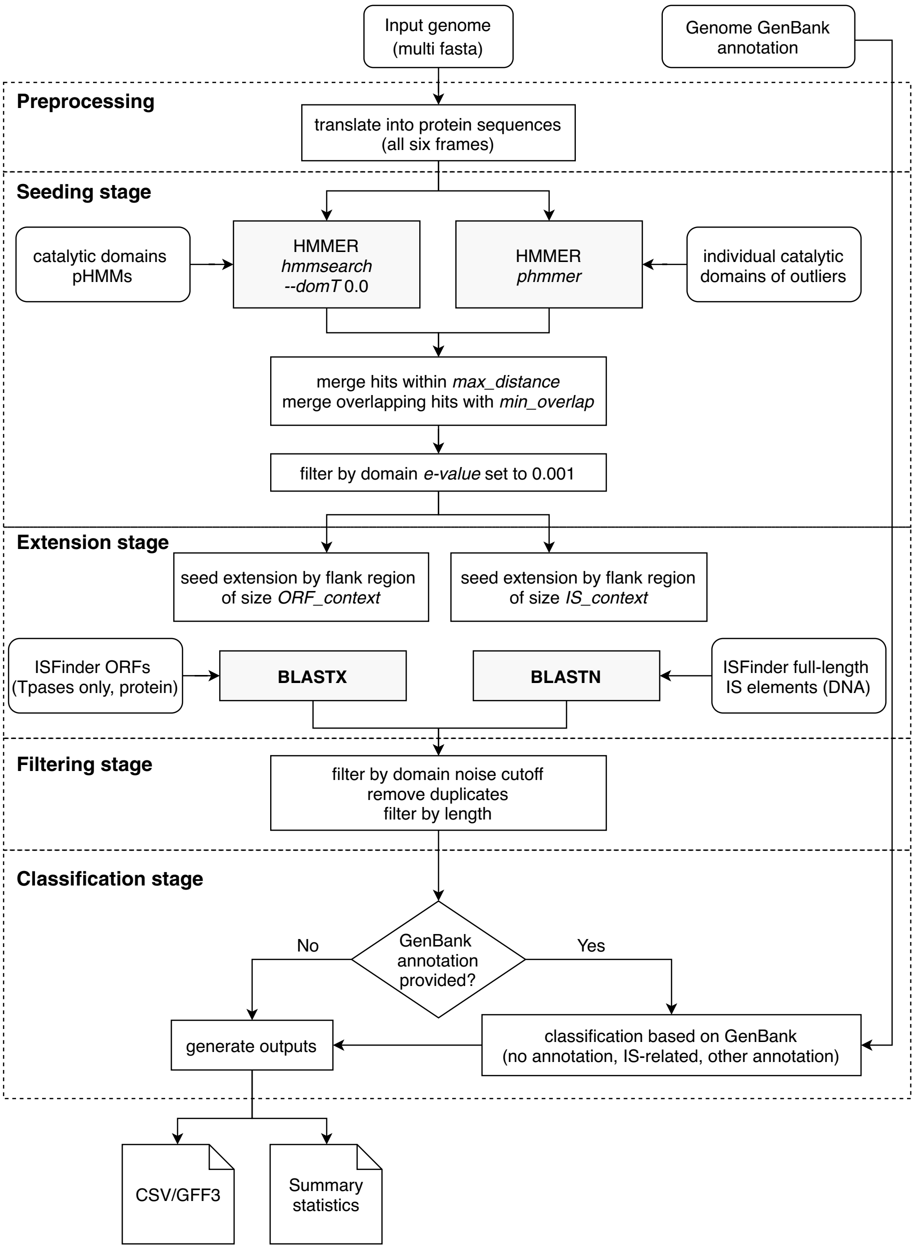

Input genome
(multi fasta)
Genome GenBank annotation
Preprocessing
translate into protein sequences (all six frames)
Seeding stage
catalytic domains pHMMs
HMMER hmmsearch --domT 0.0
HMMER phmmer
individual catalytic domains of outliers
merge hits within max_distance
merge overlapping hits with min_overlap
filter by domain e-value set to 0.001
Extension stage
seed extension by flank region of size ORF_context
seed extension by flank region of size IS_context
ISFinder ORFs (Tpases only, protein)
BLASTX
BLASTN
ISFinder full-length IS elements (DNA)
Filtering stage
filter by domain noise cutoff
remove duplicates
filter by length
Classification stage
GenBank annotation provided?
No
Yes
generate outputs
classification based on GenBank (no annotation, IS-related, other annotation)
CSV/GFF3
Summary statistics

Fig. 1 Workflow of *digIS*. digIS components and workflow, grey rectangles represent external tools, rounded rectangles represent input data, white rectangles represent digIS components

determined for each pHMM as follow: First, a database of manually curated protein sequences from Archaea and Bacteria kingdoms was collected from SwissProt [22] and RefSeq [23] databases (records labeled as 'REVIEWED'), resulting in 353051 and 232157 records (accessed on 11 March 2019), respectively. Setting this threshold is a common practice and is used, for example, in models stored in Pfam [24] database. Then, each pHMM was queried against this reference protein database employing *hmmsearch* with default settings. Finally, reported hits were sorted in a descending order based on the reported per-domain bit score and evaluated manually to estimate the bit score from which false positive hits were prevalent.

Searching for IS elements in the input sequence

In the beginning, the whole input nucleic acid sequence is translated into amino acids (all six frames). Then, the search process operates in two steps:

1 *Seeding*: The input genome is scanned using pHMMs and *individual* sequences representing Tpase catalytic domains. Each occurrence with a satisfactory score is labeled as a seed.
2 *Extension*: The genomic position of seeds identified in the previous step are extended based on the similarity boundaries with Tpases and IS elements from the ISfinder database.

In the *Seeding* stage, *digIS* utilizes *hmmsearch* from the HMMER3 package to query pHMMs against the translated sequences with an enabled domain threshold (*–domT* argument) set to 0.0 to report domain hits with a non-negative bit score only. Afterwards, *digIS* employs *phmmer* to query *individual* protein sequences against the translated sequences. The resulting hits are post-processed and filtered by a domain conditional e-value set to 0.001. Next, neighboring records, detected by the same model within a certain distance (700 bp[2]) on the same strand, are merged. This approach allows insertions or variable segments inside catalytic domains that are typical for some Tpases [25]. Next, overlapping records found by different models are merged, since there exists a sequence similarity in the catalytic domain among different Tpases, or a putative novel catalytic domain might be composed of different parts of known domains.

Please note that *digIS* scans the whole input sequence, instead of just open reading frames (ORFs), to not omit some coding regions.

During the next stage (*Extension*), the genomic position of each seed is identified in the original nucleic acid sequence and extended with *context_orf* and *context_dna* (upstream and downstream flank regions of a length 1600 bp[3], and 14000 bp[4], respectively), see Fig. 2. Next, the extended seed is matched against sequences of known Tpases (ORF level) and IS elements (DNA level), extracted from the ISfinder database, using the BLASTX and BLASTN tools. Finally, the seed's original position is adjusted (extended) according to the best BLAST hits' positions.

As the output of the *Extension* stage, the *digIS* tool reports: (1) position at DNA level if the similarity with a known IS element was found using the BLASTN tool; or (2) position at the ORF level if the similarity with a known Tpase was found using the BLASTX tool; or (3) position of the catalytic domain otherwise found during the *Seeding* stage.

[2] Merge distance 700 bp was identified based on the longest gaps within the models (see Additional file 3 for more details).

[3] ORF context size 1600 bp was identified based on the length of the longest transposase ORF in the ISfinder database divided by 2, multiplied by 3 (conversion from amino acids to nucleotides) and rounded up to the nearest hundredth

[4] DNA context size 14000 bp was identified based on the length of the longest IS element in the ISfinder database divided by 2 and rounded up to the nearest hundredth

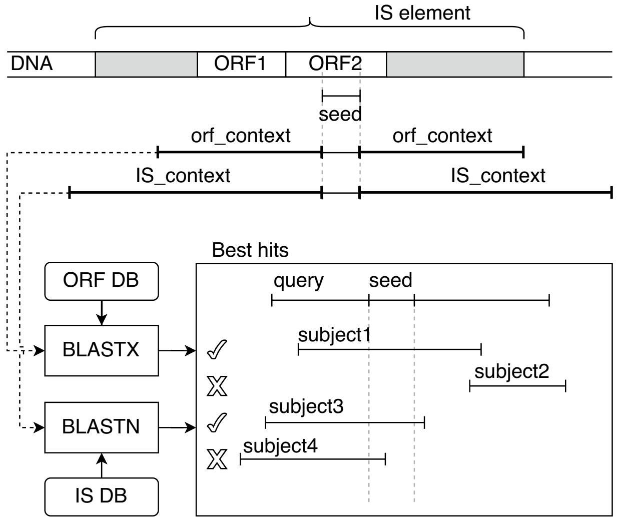


Fig. 2 Seed extension process. The seed matches the catalytic domain of the putative IS element in the input genomic sequence. This seed is extended with upstream and downstream flank regions of orf_context and IS_context size, respectively, and is searched against the database of known Tpase/ORFs and IS elements, respectively. Only the best hits, including the whole original seed, are considered for extension. Position of the seed is changed (extended) according to the best hit

Output filtering

To eliminate the number of reported false positives, *digIS* filters the hits with a score below the previously estimated noise cutoff threshold, and it removes duplicate records covering the same genomic region. Lastly, hits having less than 150 bp (50 aa) in length are filtered out.

Output classification

To help the user assess the quality of found IS elements, each output hit is supplemented by information about sequence similarity with known IS elements and Tpases extracted from the ISfinder database. The similarity is calculated as a percentage of identity between the extended seed and a known IS element or Tpase sequence, measured according to the database item's length.

In case the GenBank annotation is provided as an optional input[5], the classification process is further extended, and each *digIS* hit is classified based on the overlap with GenBank annotation records into the three categories using following rules applied in the subsequent order:

- *IS-related*—hit overlaps with a GenBank record of type: (1) mobile element or mobile element type, (2) repeat region, coding sequence (CDS), gene, or miscellaneous feature annotated as transposase, resolvase, recombinase, recombination/resolution, insertion element, mobile element, transposon, transposable element, DDE, or the

[5] GenBank annotation is a result of a complex process [26] that utilizes sources of manually curated data and automatically predicted ones with a high level of confidence.

annotation contains a name of known IS family or subfamily [27, 28]. A hit classified into this category has high confidence to be a true IS element.

- *no annotation*—hit does not overlap with any GenBank record or overlaps with a record annotated as a hypothetical protein, predicted protein, unknown, or domain of the unknown function (DUF). The hit in this category can be seen as an unknown protein or protein, where the annotation pipeline did not achieve a sufficient level of confidence. Typically, distant or putative novel IS family members may belong to this category.
- *other annotation*—otherwise. The hit in this category is probably not an IS element, because it overlaps and shares significant similarity with a different protein.

Since the previous analysis of GenBank annotation revealed that some IS element transposases were misannotated as integrases [6, 12], we classify all hits annotated as integrases and at the same time having significant identity to a known IS element in the ISfinder database (at ORF or DNA level), as *IS-related* as well.

The latest version of the GenBank annotation was newly expanded to include fragments of IS elements marked as ’pseudo’ with the notation ’incomplete’ [26]. To preserve a conservative approach and high confidence, these records are ignored when classifying hits.

digIS output files

The *digIS* tool generates the following output files: (1) a CSV and GFF3 file containing all found IS elements and their attributes such as sequence ID, genomic location, strand, accuracy, score, sequence similarities with known IS elements (at ORF and DNA level), and classification according to GenBank annotation (if provided); (2) a summary file containing numbers of IS elements per individual families, overall numbers of base pairs and a percentage of an input sequence occupied by IS elements. FASTA sequences of found IS elements can be extracted using the GFF3 file and BEDTools [29] (see instructions on the GitHub repository).

Results

The performance of the *digIS* tool was evaluated on different datasets and compared with related tools. Specifically, we chose ISEScan (version 1.6), OASIS (version released 18th September 2012), and ISsaga (version with the last update on 20th January 2020). Other state-of-the-art tools were excluded for various reasons. ISMapper, ISseeker, ISQuest, and panISa are designed for IS elements detection in raw sequence reads. TnpPred is available online only, and it is limited to protein sequences with a maximum length of 5000 amino acids. Even though the TnpPred pHMMs are available for download, it is unclear what kind of parameters or filtration mechanisms should be used during the search. Finally, we excluded IScan, because we were not able to install it, including all necessary dependencies.

All tools were run with default or recommended settings. Additionally, ISEScan was executed with two settings: (1) default configuration with the *removeShortIS* option enabled, when IS elements shorter than 400 bp or single copy IS elements without perfect

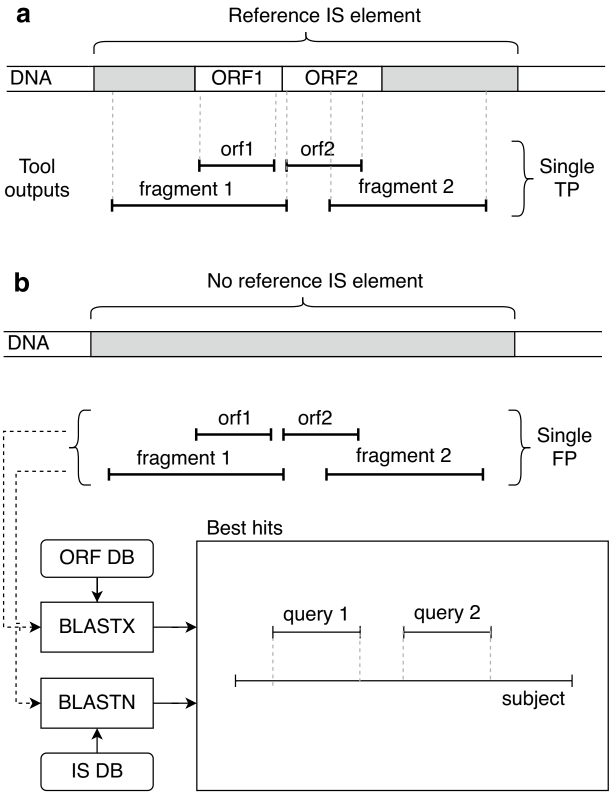


Fig. 3 Definition of true positive and false positive with respect to the tools reporting IS element's fragments or ORFs. **a** Definition of true positive: A hit is considered as a true positive (TP) if it overlaps with any item in the reference dataset. If multiple hits overlap with the same IS element in the reference dataset, then all of these hits count as one hit only. **b** Definition of false positive: Each reported hit of the tool is matched against the database of known IS elements (ISfinder), and if several adjacent hits map to the same IS element at DNA or ORF level, then these hits are counted as only one merged FP

IRs are filtered out; and (2) with *removeShortIS* turned off when all hits are reported (hereinafter referred to as ISEScan–fragments).

We faced several issues when evaluating the tools. At first, the definition of a true positive hit was ambiguous as different tools reported different types of outputs. Some tools reported entire IS elements at the DNA level (ISEScan and OASIS) or their fragments (ISEScan–fragments). Other tools reported individual ORFs or fragments thereof (ISsaga), while the proposed *digIS* tool reported outputs at one of three levels (catalytic domain, ORF, or DNA). Moreover, for tools reporting ORFs or fragments, it is common that several hits correspond to the same IS element from the reference dataset.

Considering these facts and in an effort to evaluate the tools fairly, reported hits were classified as follows: A hit is considered as a true positive (TP) if it overlaps with any item in the reference dataset, and the length of the overlapping region is $\geq$ 100 bp[6]. If multiple hits overlap with the same IS element in the reference dataset, then all of these

[6] Usually, an overlap based on a percentage of the reference IS element length is used in other studies, but when allowing for fragments, this criterion is not applicable. The requirement for at least 100 bp overlap seems to work well, even when two neighboring IS elements overlap.

hits count as one hit only (as shown in Fig. 3a). A false negative (FN) is defined as a reference dataset element without sufficient overlap with at least one reported hit. A false positive (FP) represents a reported hit without sufficient overlap with at least one item from the reference dataset.

It turns out that some reference datasets may not be complete. For example, if a new IS element is discovered, it is not included in a previously published dataset. A hit matching this new IS element is considered as an FP, even if it was identified correctly by the tool (see "Evaluation on the benchmark ISbrowser and E. coli datasets" section). The number of FPs is then even higher for tools reporting ORFs or fragments of the same IS element. To minimize this side effect, each FP was compared with a database of known IS elements (ISfinder). If several adjacent FPs mapped to the same IS element at DNA or ORF level (as shown in Fig. 3b), they were counted as one merged FP (mFP).

Evaluation on the benchmark ISbrowser and *E.coli* datasets

The first evaluation of the selected tools was performed on two benchmark datasets (1) a human-curated dataset from ISbrowser, and (2) the IS element annotation of *Escherichia coli* strain K-12 substr. MG 1655 genome [30]. The ISbrowser dataset comprises an annotation of 118 prokaryotic genomes (as of November 2017); 58 of them contain full-length IS elements, including 36 prokaryotic genomes and 22 plasmids. *E.coli* strain K-12 is one of the most well-understood model organisms [31] and is frequently used in microbial studies. The dataset of annotated IS elements for *E.coli* was obtained from the ISEScan publication (Supplementary Materials, Table 5) since EcoGene 3.0 [31], a source devoted to the structural and functional annotation of *E.coli* strain K-12, was unavailable at the time of manuscript preparation. This dataset consists of 49 IS elements of which 40 are full-length.

Results for the ISbrowser and *E.coli* datasets are shown in Table 2. Surprisingly, all tools showed a relatively high number of FPs and corresponding FDR (in the range from 8 to 24%). Therefore, we analyzed the FPs in more detail as follows: First, FPs representing fragments/ORFs of the same IS element were merged as described at the beginning of this section. Then, for each merged FP (mFP), the similarity with known IS elements in the ISfinder database was measured and by using the GenBank annotation it was classified into *IS-related*, *no annotation*, or *other annotation* category as described in "Output classification" section. Based on these results, a histogram was plotted depicting the number of mFPs as a function of similarity at both ORF and DNA levels. Finally, each bar in the histogram was divided according to the classification based on the GenBank annotation. These histograms represent an effective way to visualize the outputs of individual tools, including the identification of areas in which the tool makes errors. Please, see Additional file 4: "ISbrowser dataset" section.

In summary, many mFPs correspond to the hits that are highly likely to represent true IS elements that are not yet included in manually curated datasets. This behavior can be caused by the fact that the human-curated, whole-genome annotation might not be updated as often as databases of known IS elements. The exact numbers of true IS elements are unknown even in human-curated datasets and may evolve over time. Therefore, the common performance metrics, such as the confusion matrix, can not evaluate the tool quality fairly.

Table 1 Thresholds for classification based on the sequence similarity

Level/interpretation	Improbable member	Inter-family member	Intra-family member
IS element	SeqID < 50%	50% < SeqID ≤ 70%	70% < SeqID
Tpase/ORF	SeqID < 25%	25% < SeqID ≤ 45%	45% < SeqID

To address this issue, we decided to classify mFP hits further to distinguish between those representing IS elements with a high level of evidence and improbable/not IS elements. For these purposes, we used the GenBank annotation, which resulted from a conservative approach combining manually curated data and automatically predicted ones with a high level of confidence. Each mFP hit was classified according to the rules described in the "Output classification" section. Therefore, mFPs classified as *IS-related* can be highly likely considered as IS elements or their parts. Similarly, mFPs classified as *other annotation* can be regarded as improbable or not IS elements since they include parts that have been conservatively identified as other protein products.

The remaining hits classified as *no annotation* can be seen as unknown IS elements or those where the GenBank annotation pipeline has not achieved a sufficient level of confidence. To evaluate these outputs, additional information about sequence similarity with the database of known sequences (ISfinder) was used. Since the IS elements are divided into several independent families, it is difficult to find the exact boundary between IS and non-IS elements for mFPs. It is more appropriate to divide them into three categories:

- *Intra-family member*—a hit having similarity to the extent that is typical for members belonging to the same family.
- *Inter-family member*—a hit having similarity that is common among members of different families.
- *Improbable member*—a hit having similarity lower than usual among family members.

Although there may be several ways to categorize mFPs into these groups, we have chosen a more straightforward approach by defining two similarity thresholds (at the ORF and DNA level) that divide hits into these three categories. To determine the thresholds, a database of known IS elements (ISfinder) was used, the sequence similarities common within existing families and among them were measured, and these values were averaged. The resulting thresholds and their interpretations are given in Table 1. A detailed description of the procedure and the measured data is available in Additional file 5.

In summary, using the GenBank annotation and sequence similarity, the mFPs were classified into three categories according to the following rules:

- *IS element with a high level of evidence (eIS)*—a hit classified as *IS-related* based on the GenBank annotation, or a hit classified as *no annotation* based on the GenBank and *Intra-family member* based on the sequence similarity.
- *Distant or putative novel IS element (pNov)*—a hit classified as *no annotation* based on the GenBank and *Inter-family member* based on the sequence similarity.

- *Improbable or not an IS element (nIS)*—a hit classified as *other annotation* based on the GenBank annotation, or a hit classified as *no annotation* based on the GenBank and *Improbable member* based on the sequence similarity.

Distribution of mFP entries into these three categories is presented in Table 3, columns labeled as *Detailed classification of mFPs*. It can be seen that a large part of the hits initially classified as mFPs falls into the category *IS element with a high level of evidence*. Together with previously identified TPs, they represent the total number of IS elements with a high level of evidence (teIS). Consequently, only the hits in the nIS category are considered to be incorrectly identified by the tool (i.e. false positives). Based on these new metrics, the putative novel discovery rate (pNovDR), and nIS discovery rate (nISDR) were calculated representing the proportion of putative novel and improbable/not IS elements in reported outputs, respectively. Finally, the pNov/nIS ratio was calculated to express how many putative novel elements are found per single incorrectly identified hit.

We presume that these modified metrics reflect the tools' performance better since they address the issue of incomplete reference datasets. Concurrently, they are based on sequence similarity information with known IS elements (ISfinder) and state-of-the-art annotations with high confidence (GenBank). We are aware of possible discussions and alternatives towards defined classification rules and similarity thresholds. However, if they are applied to all tools equally, they can bring a more reliable image of their performance.

The results in Table 3 related to the ISbrowser dataset show that:

- The tools that detect both full elements and fragments (ISsaga and ISEScan–fragments) can find the highest number of teISs. On the other hand, the reported hits include the highest number of nISs. The overall nISDR is around 9%, and the ratio between pNovs and nISs is low (0.15 and 0.22).
- OASIS found the lowest number of teISs and nISs (nISDR is 1.15%), making it the most conservative tool of all. OASIS found only the hits with a high level of confidence. The output primarily includes records of known IS elements, whereas putative novel elements are rare (0.69%).
- ISEScan is the second most conservative tool in terms of the number of teISs and nISs. Surprisingly, it found even less pNovs compared to the OASIS tool.
- With respect to the number of teISs and nISs, *digIS* falls in the middle between conservative (OASIS and ISEScan) and fragment-reporting tools (ISEScan–fragments and ISsaga) representing a tool with good sensitivity (0.82) and low nISDR (3.58%). Moreover, the number of pNovs is even higher than for ISEScan–fragments. Although ISsaga found one-third more pNovs than *digIS*, it was at the cost of three times more nISs.

The tools show a similar performance on the *E.coli* dataset. However, some characteristics are violated; for instance, none of the tools found any putative novel element, and nISDR is more than double for most tools. These discrepancies are primarily caused by a too small *E.coli* dataset (a single genome with less than 50 IS elements), where some of the metrics are calculated from fewer than ten items. Similar distortion can also be seen in the ISbrowser dataset, where the numbers of pNovs and nISs are too small for the OASIS tool. It results in a disproportionately high pNov/nIS ratio.

Evaluation on the NCBI Archaea and Bacteria datasets

In the next step, tools were evaluated on much larger datasets to verify the characteristics observed in Table 3 and to specify those affected by the small number of samples. We prepared two additional datasets containing complete archaeal and bacterial genomes from the NCBI GenBank database [32]. In the case of Archaea, all 341 genomes available in the database were used (accessed on 15th June 2019). In the case of Bacteria, 2500 from 14418 available genomes were randomly selected (see Additional file 6 for detailed information about these datasets). Since OASIS could not process 25 bacterial genomes, these were excluded. Altogether, 2475 bacterial genomes were evaluated.

Unlike the ISbrowser and *E.coli* datasets, the manually curated positions of IS elements are not available. Therefore, all hits reported by the tools were considered as FPs and the detailed classification process of FPs described in “Evaluation on the benchmark ISbrowser and E. coli datasets” section was applied. To verify the accuracy of this evaluation method, it was applied to the ISbrowser dataset first. Table 4 shows the number of hits found by the tool (N), the number of merged FPs (mFPs), the output of the classification process (number of eISs, pNovs, and nISs), and an assessment in terms of pNovDR, nISDR, and pNov/nIS ratio. As the number of TPs is not available, the teIS is reduced to eIS.

By comparing the evaluation results for the ISbrowser dataset with and without human-curated annotation (Tables 3, 4), certain differences can be seen. Detailed analysis revealed that these changes arose primarily because the ISbrowser reference dataset contains not only full-length elements, but also annotated fragments of various lengths (a total of 127 fragments). If a tool finds some of these fragments, they are distributed among the categories eIS, pNov, and nIS based on the GenBank annotations and similarities with the ISfinder database. This behavior causes the number of pNovs and nISs to increase at the expense of the total number of eIS. As a side effect, the pNovDR, nISDR, and pNov/nIS ratio are slightly higher. The small changes can also be observed in the histograms (see Additional file 4: “ISbrowser dataset without reference” section), but their overall character remains the same. Considering these subtle differences, it is possible to conclude that the above-described classification allows us an assessment of the tool performance, even when the manually curated annotation is not available.

The results on large NCBI GenBank Archaea and Bacteria datasets in Table 4 confirmed the tools’ characteristics seen on the ISbrowser dataset. Only the following differences were observed:

- The proportion of nISs in the outputs is higher compared to the ISbrowser dataset. For ISEScan and *digIS*, the nISDR is approximately twice as large on the Archaea dataset. ISsaga achieved the highest nISDR (around 20%) for both Archaea and Bacteria datasets. A detailed analysis of the hits revealed that this is primarily due to the higher number of items classified as *other annotation*. A list of the most common GenBank record products that overlapped with these hits is given in Additional file 7.

- Larger NCBI datasets enabled to assess the ratio between pNov and nIS for OASIS more accurately, as it was affected by a small number of items in the *E.coli* and ISbrowser datasets before. This ratio decreased significantly to 0.27 and 0.21. Also, the number of pNovs found by OASIS is no longer higher than those found by the ISEScan tool.
- The histograms depicting the similarity of the outputs with the ISfinder database and their classification according to the GenBank annotation show the same characteristics as for the ISbrowser dataset, except for minor deviations (see Additional file 4: "NCBI Archaea and Bacteria datasets without reference" section).

In summary, tools that also detect fragments (ISsaga and ISEScan–fragments) can identify the most eISs, but at the cost of a large number of nISs. On the other side of the spectrum are conservative tools (OASIS and ISEScan), which show the lowest numbers of nISs, but also eISs. The performance of the proposed *digIS* tool in terms of eISs is closer to fragment-reporting tools, and at the same time, it achieves the number of nISs closer to conservative tools. Moreover, *digIS* is dominant in finding distant/putative novel IS elements with respect to the numbers of nISs (pNov/nIS ratio). This feature is significant, especially on large datasets (NCBI GenBank Archaea/Bacteria), where the *digIS* tool shows the best performance. Please note that *digIS* found even more putative novel elements than the ISEScan–fragments in these datasets.

Discussion

In this work, we focused on the detection of putative novel IS elements and aimed to find the sequence and structural features common to more IS families. The Tpases are generally considered as the most conserved parts of IS elements. Their structural variability is used as a major feature for their classification into the families [1]. On the other hand, the Tpase catalytic domain and its secondary structure are often preserved among the families [25]. Unfortunately, the accuracy of state-of-the-art tools for secondary structure prediction is not sufficient when applied to a single sequence and MSA is usually required for a more accurate prediction [33].

For this reason, we decided to make a compromise between detecting the general structure and sequence features. We built the library of manually curated pHMMs of a catalytic domain only (not whole transposase). The results of comparing *digIS* with other tools confirmed that the search based on the catalytic domain is sufficiently specific for the area of IS elements. The number of IS elements with a high level of evidence is comparable to fragment-reporting tools, while many improbable/not IS elements are filtered out. To better understand the effectiveness of the catalytic-domain-search technique compared to using the pHMM of the whole Tpase sequence, we performed a detailed analysis of individual tools' outputs. We focused on hits classified as "other annotation" according to the GenBank annotation, i.e., the records erroneously identified by the tool as IS elements or their parts. We analyzed overlapping GenBank records for these hits and created a histogram showing the number of occurrences for each type of protein or product (see Additional file 7).

Table 2 Performance of OASIS, ISEScan, ISEScan-fragments, ISsaga, and digIS on manually curated datasets

Tool	TP	FN	FP	Se	FDR (%)
Dataset ISbrowser (N = 1192)					
OASIS	791	401	77	0.66	8.87
ISEScan	925	267	94	0.78	9.22
ISEScan-fragments	1077	115	248	0.90	18.71
ISsaga	1135	57	363	0.95	24.23
digIS	979	213	194	0.82	16.54
Dataset *E. coli* (N = 49)					
OASIS	26	23	4	0.53	13.33
ISEScan	43	6	8	0.88	15.69
ISEScan-fragments	45	4	18	0.92	28.57
ISsaga	48	1	29	0.98	37.66
digIS	43	6	11	0.88	20.37

TP, FN, and FP represent the number of True Positives, False Negatives, and False Positives, respectively; Se is sensitivity; FDR is False Discovery Rate.

Table 3 Detailed analysis of false positives of digIS, ISEScan, OASIS, and ISsaga on manually curated

Tool	Common metrics		Detailed classification of mFPs				Modified metrics			
	TP	FP	mFP	eIS	pNov	nIS	teIS	pNovDR (%)	nISDR (%)	pNov/nIS
Dataset ISbrowser (N = 1192)										
OASIS	791	77	75	59	6	10	850	0.69	1.15	0.60
ISEScan	925	94	94	69	3	22	993	0.29	2.16	0.14
ISEScan-fragments	1077	248	239	103	18	118	1179	1.37	8.97	0.15
ISsaga	1135	363	323	148	31	144	1282	2.13	9.88	0.22
digIS	979	194	194	130	22	42	1108	1.88	3.58	0.52
Dataset *E. coli* (N = 49)										
OASIS	26	4	4	4	0	0	30	0.00	0.00	0.00
ISEScan	43	8	7	3	0	4	46	0.00	8.00	0.00
ISEScan-fragments	45	18	17	6	0	11	51	0.00	17.74	0.00
ISsaga	48	29	28	10	0	18	58	0.00	23.68	0.00
digIS	43	11	11	6	0	5	49	0.00	9.26	0.00

N represents the number of outputs found by the tool; mFP represents the number of False Positives after merging fragments or ORFs referencing the same IS element; eIS, pNov, and nIS represent the number of mFPs classified into categories IS element with a high level of evidence, Distant or putative novel IS element, and Improbable or not an IS element, respectively; pNovDR is putative Novel Discovery Rate; nISDR is Improbable or not an IS element Discovery Rate, and pNov/nIS shows the ratio between the number of putative novel IS elements and improbable or not an IS elements.

From the generated histograms, it can be observed that *digIS* generally reports a small number of records classified as "other annotation", which is comparable to conservative tools such as OASIS or ISEScan (see Additional file 7; Tables 1, 2, 3). On the other hand, tools that also report fragments (ISsaga and ISEScan–fragments) show a large number of these hits. If we focus on the annotations of these records, it can be seen that they usually represent products functionally related to transposases or parts thereof, such as *DNA-binding protein*, *ATP-binding protein*, *transcriptional regulator*, or *helix-turn-helix domain-containing protein*. In addition, both fragment-reporting tools (ISsaga and ISEScan–fragments) cover a large number of products that were not observed by other tools,

Table 4 Performance of digIS against ISEScan, OASIS, and ISsaga on NCBI GenBank datasets

Tool	N	Detailed classification of mFPs				Modified metrics		
		mFP	eIS	pNov	nIS	pNovDR (%)	nISDR (%)	pNov/nIS
Dataset ISbrowser (N = 1192)								
OASIS	895	852	828	10	14	1.17	1.64	0.71
ISEScan	1006	993	954	9	30	0.91	3.02	0.30
ISEScan-fragments	1326	1283	1089	41	153	3.20	11.93	0.27
ISsaga	1786	1459	1188	75	196	5.14	13.43	0.38
digIS	1170	1157	1051	50	56	4.32	4.84	0.89
Dataset NCBI Archaea (341 genomes)								
OASIS	5885	5789	5382	100	307	1.73	5.30	0.33
ISEScan	8404	8266	7532	207	527	2.50	6.38	0.39
ISEScan-fragments	12,016	11,550	9622	472	1456	4.09	12.61	0.32
ISsaga	17,698	14,788	10,946	822	3020	5.56	20.42	0.27
digIS	10,607	10,548	8640	728	1180	6.90	11.19	0.62
Dataset NCBI Bacteria (random selection of 2475 genomes)								
OASIS	88,552	87,428	83,992	1176	2260	1.35	2.58	0.52
ISEScan	111,974	110,357	102,266	3274	4817	2.97	4.36	0.58
ISEScan-fragments	151,540	145,248	119,392	6096	19760	4.20	13.60	0.31
ISsaga	217,345	181,880	136,903	8479	36,498	4.66	20.07	0.23
digIS	134,851	132,877	118,805	6722	7350	5.06	5.53	0.91

N represents the number of outputs found by the tool; mFP represents the number of False Positives after merging fragments or ORFs referencing the same IS element; eIS, pNov, and nIS represent the number of mFPs classified into categories IS element with a high level of evidence, Distant or putative novel IS element, and Improbable or not an IS element, respectively; pNovDR is putative Novel Discovery Rate; nISDR is Improbable or not an IS element Discovery Rate, and pNov/nIS shows the ratio between the number of putative novel IS elements and improbable or not an IS elements.

including *digIS*, such as *chromosomal replication initiator protein DnaA*, *DNA replication protein DnaC*, or *primosomal protein DnaI*. Detailed analysis revealed that portions of these proteins have significant sequence similarity to the coding segments of IS elements of the IS21 family (see Additional file 7). These examples show that searching for any fragments of IS elements can lead to a large number of false hits, which the application user must manually check. On the other hand, focusing the search on the catalytic domain can effectively filter these hits and, unlike conservative methods reporting full-length elements only, it provides a space for searching for putative novel IS elements.

When comparing the tools without a manually curated reference dataset or an incomplete one, the histogram—showing the number of outputs depending on the similarity to the database of known elements (ISfinder) and GenBank annotation—is a useful indicator of the tool's quality. It offers an independent view of the characteristics of the outputs and clearly shows, for example, the degree of tool conservation or tendency to detect other genes, that is typical for fragment-reporting tools (ISsaga and ISEScan–fragments). It also allows the identification of various anomalies in the GenBank annotation itself (see Additional file 4).

Despite the histogram's benefits, it does not allow us to easily quantify and compare the performance of the tools. The comparison is possible only if the outputs are classified into distinct categories such as TPs, TNs, FPs, FNs using manually curated benchmark datasets. In this paper, we were the first to point out the drawbacks of this approach when applied to existing tools for IS elements detection. We addressed the issue of different outputs of individual tools (full-length elements vs. fragments/ORFs). Based on a detailed analysis (see Additional file 4), we have shown that the benchmark datasets themselves are not complete, and therefore their use may skew the evaluation results.

To overcome these issues, we have chosen an alternative classification of the tools' outputs that relies on GenBank annotation and sequence similarity with the database of known elements (ISfinder). This approach allowed us to identify a group of IS elements with a high level of evidence (eIS) and a group of Improbable or not IS elements (nIS) in the category of presumed false positives. Also, since the boundary between these two groups is not strictly defined, there is a space for the putative novel IS elements group (pNov), which is the main interest of this article. We are aware that the definition of these categories is unambiguous and should be replaced by a high-quality and consistently maintained benchmark dataset in the future. On the other hand, the boundary between the groups of pNovs and nISs will probably be the subject of debate for a long time, as its precise definition would require a knowledge of all non-IS elements.

We experimented, for example, with a different definition of pNov and its effect on tools performance. Currently, pNov is defined as a sequence without a sufficiently specific GenBank annotation, having the sequence similarity that is common among members of different IS families. Without further restrictions, this category may include, for example, the found accessory genes or some of the transposase's variable domains. To make sure that the found hit is highly likely functional from a transposition point of view, it would be appropriate to require the presence of Tpase and its catalytic domain. Therefore, an analysis of the pNov hits was performed and those that overlap with the catalytic domain of any known IS element were identified (see Additional file 8). This analysis showed that many hits fall outside the catalytic domain, especially for fragment-reporting tools (ISsaga and ISEScan−fragments). If the tools were evaluated according to this stricter definition, then the proposed *digIS* would achieve the best results in the detection of pNovs on an absolute scale.

We analyzed the coverage of pNovs by individual tools to identify which of them are reported by several tools simultaneously or, conversely, exclusively by a specific tool. We also measured pNovs regarding their proximity to existing families of IS elements to reveal a possible preference of the tool to search for pNovs in a certain part of the sequence space (see Additional file 8). It turned out that various tools have a preference to search pNov elements close to various IS families. For example, *digIS* found the most pNovs close to the ISH3 family while ISsaga found the most pNovs close to the IS5 family. In summary, it can be concluded that no tool would include all pNov outputs of other tools.

Finally, we performed an analysis of the found pNovs to verify that they met the common characteristics of IS elements, such as multiple occurrences in the genome, or the presence of IR and DR regions. Using clustering, we found groups of similar hits, then performed their multiple sequence alignment, and identified IR and DR regions. Based on a manual inspection of selected clusters, we have identified four novel IS elements, of which the first two can be found by competing tools and the other two represent new ones found exclusively by the *digIS* tool (see Additional file 9).

Conclusions

In this paper we present a novel approach for IS elements detection, that is implemented in the form of *digIS* tool. It combines searching for the catalytic domains of transposases and additional filtering mechanisms that allows to detect not only known IS elements, but also distant putative novel IS elements. Simultaneously, it eliminates a large number of false hits that are typical for fragment–reporting tools.

Comparison with other state-of-the-art tools, such as ISsaga, OASIS, and ISEScan, on different datasets (*E.coli*, ISbrowser, NCBI GenBank Archaea/Bacteria) confirmed that *digIS* can find the majority of known ISs and shows the best ratio between putative novel elements and improbable/not IS elements. This makes it the right choice for scientists who are interested in finding new IS elements.

Finally, we would also highlight the technical aspects of the developed software. *digIS* is one of the few tools that still works and is ready for future use in the form of a Docker image. Simultaneously, it does not limit the user in the number of sequences to be analyzed or other search parameters, as is the case of web-based tools. *digIS* is ready to run in a grid-computing and cloud environment, which is very important for scalability. The transparency and credibility of the tool are further supported by the open-source code on GitHub (Table 4).

Abbreviations

bp:: Base pair; CDS:: Coding sequence; DR:: Direct repeat; DUF:: Domain of unknown function; eIS:: IS element with a high level of evidence; FN:: False negative; FP:: False positive; IR:: Inverted repeat; IS:: Insertion sequence; kbp:: Kilobase pair; MSA:: Multiple sequence alignment; mFP:: Merged false positive; nIS:: Improbable or not an IS element; nISDR:: Improbable/not an IS element discovery rate; ORF:: Open reading frame; pNov:: Putative novel; pNovDR:: Putative novel discovery rate; pHMM:: Profile hidden Markov model; TP:: True positive; teIS:: Total number of IS element with a high level of evidence.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12859-021-04177-6.

Additional file 1. Multiple sequence alignment of IS5/IS5 subfamily.

Additional file 2. Multiple sequence alignment of IS5/None subfamily.

Additional file 3. Multiple sequence alignment of ISL3 family.

Additional file 4. Analysis of merged FPs.

Additional file 5. Calculation of the similarity at IS and ORF level.

Additional file 6. Detailed information about NCBI GenBank archaeal and bacterial genomes used in the evaluation.

Additional file 7. Analysis of hits classified as *other annotation.*

Additional file 8. Analysis of putative novel elements.

Additional file 9. Putative novel IS elements detected by *digIS.*

Acknowledgements

Computational resources were supplied by the project "e-Infrastruktura CZ" (e-INFRA LM2018140) provided within the program Projects of Large Research, Development and Innovations Infrastructures. We would like to thank Alessandro Varani (ISsaga developer) for his help, Andrea Bisconti for English proofreading and correction, and two anonymous reviewers for their valuable feedback which significantly significantly improved the quality of this work.

Authors' contributions

JP and TM designed the research. The code was written and developed by JP and TM. Data was obtained by JP and TM. The evaluation was performed by JP and TM. All authors have read and approved the final manuscript.

Funding

This work was supported by the Ministry of Education, Youth and Sports of the Czech Republic from the National Programme of Sustainability II, IT4Innovations excellence in science [grant number LQ1602]. All data was publicly sourced

and collected from public repositories. None of the funding bodies played a role in the design of the study and collection, analysis and interpretation of data and in writing the manuscript.

Availability of data and materials

The reference genomes used during the current study are publicly available and were downloaded from the NCBI GenBank database (https://www.ncbi.nlm.nih.gov/genbank). The detailed instructions for obtaining sequences included in the NCBI Archaea and Bacteria datasets are described in Additional file 6. The manually annotated IS elements for ***E.coli*** were obtained from the ISEScan publication (Supplementary Materials, Table 5) since EcoGene 3.0 [31], a source devoted to the structural and functional annotation of ***E.coli*** strain K-12, was unavailable at the time of manuscript preparation.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Received: 30 April 2020 Accepted: 9 May 2021
Published online: 20 May 2021

References

1. Siguier P, Gourbeyre E, Chandler M. Bacterial insertion sequences: their genomic impact and diversity. FEMS Microbiol Rev. 2014;38(5):865–91. https://doi.org/10.1111/1574-6976.12067.
2. Vandecraen J, Chandler M, Aertsen A, Van Houdt R. The impact of insertion sequences on bacterial genome plasticity and adaptability. Crit Rev Microbiol. 2017;43(6):709–30. https://doi.org/10.1080/1040841X.2017.1303661.
3. Siguier PI. The reference centre for bacterial insertion sequences. Nucleic Acids Res. 2006;34(90001):32–6. https://doi.org/10.1093/nar/gkj014.
4. Kichenaradja P, Siguier P, Pérochon J, Chandler M. ISbrowser: an extension of ISfinder for visualizing insertion sequences in prokaryotic genomes. Nucleic Acids Res. 2009;38(SUPPL.1):62–8. https://doi.org/10.1093/nar/gkp947.
5. Leplae R, Lima-Mendez G, Toussaint A. ACLAME: a classification of mobile genetic elements, update 2010. Nucleic Acids Res. 2010;38(suppl–1):57–61. https://doi.org/10.1093/nar/gkp938.
6. Biswas A, Gauthier DT, Ranjan D, Zubair M. ISQuest: finding insertion sequences in prokaryotic sequence fragment data. Bioinformatics. 2015;31(21):3406–12. https://doi.org/10.1093/bioinformatics/btv388.
7. Hawkey J, et al. ISMapper: identifying transposase insertion sites in bacterial genomes from short read sequence data. BMC Genom. 2015;16(1):1–11. https://doi.org/10.1186/s12864-015-1860-2.
8. Wright MS, Bishop B, Adams MD. Quantitative assessment of insertion sequence impact on bacterial genome architecture. Microbial Genomics. 2016. https://doi.org/10.1099/mgen.0.000062.
9. Treepong P, Guyeux C, Meunier A, Couchoud C, Hocquet D, Valot B. panISa: Ab initio detection of insertion sequences in bacterial genomes from short read sequence data. Bioinformatics. 2018;34(22):3795–800. https://doi.org/10.1093/bioinformatics/bty479.
10. Wagner A, Lewis C, Bichsel M. A survey of bacterial insertion sequences using IScan. Nucleic Acids Res. 2007;35(16):5284–93. https://doi.org/10.1093/nar/gkm597.
11. Varani AM, Siguier P, Gourbeyre E, Charneau V, Chandler M. ISsaga is an ensemble of web-based methods for high throughput identification and semi-automatic annotation of insertion sequences in prokaryotic genomes. Genome Biol. 2011;12(3):30. https://doi.org/10.1186/gb-2011-12-3-r30.
12. Robinson DG, Lee M-C, Marx CJ. OASIS: an automated program for global investigation of bacterial and archaeal insertion sequences. Nucleic Acids Res. 2012;40(22):174. https://doi.org/10.1093/nar/gks778.
13. Xie Z, Tang H. ISEScan: automated identification of insertion sequence elements in prokaryotic genomes. Bioinformatics. 2017;33(21):3340–7. https://doi.org/10.1093/bioinformatics/btx433.
14. Riadi G, Medina-Moenne C, Holmes DS. TnpPred: a web service for the robust prediction of prokaryotic transposases. Comp Funct Genomics. 2012;2012:678761. https://doi.org/10.1155/2012/678761.
15. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215(3):403–10. https://doi.org/10.1016/S0022-2836(05)80360-2.
16. Delcher AL, Bratke KA, Powers EC, Salzberg SL. Identifying bacterial genes and endosymbiont DNA with Glimmer. Bioinformatics. 2007;23(6):673–9. https://doi.org/10.1093/bioinformatics/btm009.
17. Mistry J, Finn RD, Eddy SR, Bateman A, Punta M. Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. Nucleic Acids Res. 2013;41(12):121. https://doi.org/10.1093/nar/gkt263.
18. Cock PJA, et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics. 2009;25(11):1422–3. https://doi.org/10.1093/bioinformatics/btp163.
19. Sievers F, et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol Syst Biol. 2011. https://doi.org/10.1038/msb.2011.75.
20. Waterhouse AM, Procter JB, Martin DMA, Clamp M, Barton GJ. Jalview version 2-a multiple sequence alignment editor and analysis workbench. Bioinformatics. 2009;25(9):1189–91. https://doi.org/10.1093/bioinformatics/btp033.

21. Drozdetskiy A, Cole C, Procter J, Barton GJ. JPred4: a protein secondary structure prediction server. Nucleic Acids Res. 2015;43(W1):389–94. https://doi.org/10.1093/nar/gkv332.
22. Boutet E, Lieberherr D, Tognolli M, Schneider M, Bairoch A. UniProtKB/Swiss-Prot. Methods Mol Biol (Clifton, NJ). 2007;406:89–112.
23. O'Leary NA, et al. Reference sequence (refseq) database at ncbi: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res. 2016;44(D1):733–45. https://doi.org/10.1093/nar/gkv1189.
24. El-Gebali S, Mistry J, Bateman A, Eddy SR, Luciani A, Potter SC, Qureshi M, Richardson LJ, Salazar GA, Smart A, Sonnhammer ELL, Hirsh L, Paladin L, Piovesan D, Tosatto SCE, Finn RD. The Pfam protein families database in 2019. Nucleic Acids Res. 2019;47(D1):427–32. https://doi.org/10.1093/nar/gky995.
25. Majorek KA, et al. The RNase H-like superfamily: new members, comparative structural analysis and evolutionary classification. Nucleic Acids Res. 2014;42(7):4160–79. https://doi.org/10.1093/nar/gkt1414.
26. Haft DH, et al. RefSeq: an update on prokaryotic genome annotation and curation. Nucleic Acids Res. 2018;46(D1):851–60. https://doi.org/10.1093/nar/gkx1068.
27. Smith MCM, Thorpe HM. Diversity in the serine recombinases. Mol Microbiol. 2002;44(2):299–307. https://doi.org/10.1046/j.1365-2958.2002.02891.x.
28. Boocock MR, Rice PA. A proposed mechanism for IS607-family serine transposases. Mobile DNA. 2013;4(1):24. https://doi.org/10.1186/1759-8753-4-24.
29. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26(6):841–2. https://doi.org/10.1093/bioinformatics/btq033.
30. Hayashi K, et al. Highly accurate genome sequences of *Escherichia coli* K-12 strains MG1655 and W3110. Mol Syst Biol. 2006. https://doi.org/10.1038/msb4100049.
31. Zhou J, Rudd KE. EcoGene 30. Nucleic Acids Res. 2013. https://doi.org/10.1093/nar/gks1235.
32. Sayers EW, Cavanaugh M, Clark K, Ostell J, Pruitt KD, Karsch-Mizrachi I. GenBank. Nucleic Acids Res. 2019;47(D1):94–9. https://doi.org/10.1093/nar/gky989.
33. Jiang Q, Jin X, Lee S-J, Yao S. Protein secondary structure prediction: a survey of the state of the art. J Mol Graph Model. 2017;76:379–402. https://doi.org/10.1016/j.jmgm.2017.07.015.

A.11 Paper XI

Satellite DNA and Transposable Elements in Seabuckthorn (Hippophae rhamnoides), a Dioecious Plant with Small Y and Large X Chromosomes

Satellite DNA and Transposable Elements in Seabuckthorn (*Hippophae rhamnoides*), a Dioecious Plant with Small Y and Large X Chromosomes

Janka Puterova[1,2], Olga Razumova[3], Tomas Martinek[2], Oleg Alexandrov[3], Mikhail Divashuk[3], Zdenek Kubat[1], Roman Hobza[1,4], Gennady Karlov[3,5], and Eduard Kejnovsky[1,*]

[1]Department of Plant Developmental Genetics, Institute of Biophysics, Academy of Sciences of the Czech Republic, Brno, Czech Republic

[2]Department of Information Systems, Faculty of Information Technology, Brno University of Technology, Brno, Czech Republic

[3]Centre for Molecular Biotechnology, Russian State Agrarian University – Moscow Timiryazev Agricultural Academy, Moscow, Russia

[4]Institute of Experimental Botany, Center of the Region Haná for Biotechnological and Agricultural Research, Olomouc, Czech Republic

[5]All-Russia Research Institute of Agricultural Biotechnology, Moscow, Russia

*Corresponding author: E-mail: kejnovsk@ibp.cz.

Accepted: January 3, 2017

Data deposition: This project has been deposited in European Nucleotide Archive (ENA) under study accession PRJEB14480. http://www.ebi.ac.uk/ena/data/view/PRJEB14480.

Abstract

Seabuckthorn (*Hippophae rhamnoides*) is a dioecious shrub commonly used in the pharmaceutical, cosmetic, and environmental industry as a source of oil, minerals and vitamins. In this study, we analyzed the transposable elements and satellites in its genome. We carried out Illumina DNA sequencing and reconstructed the main repetitive DNA sequences. For data analysis, we developed a new bioinformatics approach for advanced satellite DNA analysis and showed that about 25% of the genome consists of satellite DNA and about 24% is formed of transposable elements, dominated by Ty3/*Gypsy* and Ty1/*Copia* LTR retrotransposons. FISH mapping revealed X chromosome-accumulated, Y chromosome-specific or both sex chromosomes-accumulated satellites but most satellites were found on autosomes. Transposable elements were located mostly in the subtelomeres of all chromosomes. The 5S rDNA and 45S rDNA were localized on one autosomal locus each. Although we demonstrated the small size of the Y chromosome of the seabuckthorn and accumulated satellite DNA there, we were unable to estimate the age and extent of the Y chromosome degeneration. Analysis of dioecious relatives such as Shepherdia would shed more light on the evolution of these sex chromosomes.

Key words: sex chromosomes, genome composition, chromosomal localization, repetitive DNA.

Introduction

Seabuckthorn (*Hippophae rhamnoides*) is a hardy, deciduous dioecious shrub belonging to the Elaeagnaceae family with a natural habitat extending widely across Europe and Asia. It is used in traditional Chinese, Tibetan and Siberian medicine and has special characteristics exploitable in biotechnology, pharmaceutical and cosmetic sciences, as a source of oil, minerals and vitamins. The size of seabuckthorn genome is ~2.55 Gbp/2C (Zhou et al. 2010) but there is a dearth of information on its composition. The ribosomal DNA ITS regions were compared among *H. rhamnoides* ssp chinensis from different geographical areas of China and showed distinct genetic variation (Chen et al 2010). RAPD markers (Sharma et al. 2010) were identified with the aim of determining the sex of individuals. Cytogenetic analysis is represented only by the older works of Shchapov (1979) and Rousi and Arohonka (1980) who both determined the diploid chromosome number $2n = 24$. Shchapov (1979) revealed the small Y and large X chromosomes. Seabuckthorn transcriptome has been analyzed recently providing a resource for gene discovery and development of molecular markers (Ghangal et al. 2013).

Sex chromosomes have evolved repeatedly and independently in the plant kingdom with different age and degree of degeneration shown in various dioecious species (Ming et al.

2011; Hobza and Vyskot 2015; Charlesworth 2016). The evolution of the Y chromosomes is characterized by gene erosion/ loss and accumulation of repetitive DNA (Kejnovsky et al. 2009). The most studied dioecious model species with heteromorphic sex chromosomes are white campion (*Silene latifolia*, Kejnovsky and Vyskot 2010), sorrel (*Rumex acetosa*, Steflova et al. 2013; *R. hastatulus*, Hough et al. 2014), ivy gourd (*Coccinia grandis*, Sousa et al. 2013), and members of the Cannabaceae family (*Humulus lupulus*, Divashuk et al. 2011; *H. japonicus*, Alexandrov et al. 2012; *Cannabis sativa*, Divashuk et al. 2014).

The majority of large plant genomes are formed of repetitive DNA, mostly by transposable elements and tandem repeats (satellite DNA). The processes of repetitive DNA amplification and elimination are only partially understood. Turnover of repeats is high and corresponds only to million of years (Lim et al. 2007). The localization of repetitive DNA on sex chromosomes is different from that of autosomes, reflecting different repeat dynamics, especially on the nonrecombining regions of the Y chromosomes (Kejnovsky et al. 2009). Satellite DNA has mostly discrete localization in the genome and some satellites are thus Y chromosome-specific (Mariotti et al. 2009). In contrast, transposable elements have more homogenous distribution and are only slightly enriched on the Y chromosome (Charlesworth 1991; Cermak et al. 2008) or alternatively absent on the Y chromosome as shown in *Silene latifolia* (Cermak et al. 2008; Kubat et al. 2014) and *Rumex acetosa* (Steflova et al. 2013) despite their presence in the rest of genome. The striking example is the large Y chromosome of the dioecious plant *Coccinia grandis* showing accumulation of transposable elements, satellites, and organellar DNA (Souza et al. 2016). One review published recently discusses the role of repetitive DNA in the evolution of sex chromosomes and includes a database of transposable elements of dioecious plants (Li et al. 2016a, 2016b).

In this study, we analyzed the transposable elements and satellites in the seabuckthorn genome and determined the chromosomal localization of these repeats. We showed that seabuckthorn has an XY system with large X and small Y chromosomes.

Materials and Methods

Illumina Sequencing

DNA isolation from male (Pollinator 1) and female (cv "Botanicheskaya lyubitelskaya") plants was carried out according to Doyle and Doyle (1990). One Illumina MiSeq sequencing run was performed for each male and female genomic DNA. The voucher specimen of the plants used in the study was kept for record in the herbarium (AT) of Department of Botany and Breeding of Horticultural Crops of the Russian State Agrarian University – MTAA (Voucher No.5470). Sequencing reads were analyzed by quality control tool FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/; last accessed January 4, 2017) followed by quality filtering based on the sequence quality score, adaptors trimming, filtering out short or unpaired sequences and trimming all reads to lengths of 230 nucleotides using the Trimmomatic tool (Bolger et al. 2014), leading to 1,848,543 male and 1,863,670 female paired-end reads. Quality-filtered reads were randomly sampled to 415,650 paired-end reads for both male and female individuals and the reads were merged together (totally 1,662,600 reads). As the nuclear DNA content of *H. rhamnoides* reported in Zhou et al. (2010) was determined to be ~2.61/2C pg (without detailed specification of male or female) we converted it to genome size (in bp) using following formula (Doležel et al. 2003): $g = \text{DNA content (pg)} \times (0.978 \times 10^9)$, resulting into ~2.55 Gbp/2C, our samples represent ~30% of haploid genome. Genome coverage was calculated as follow: $cov = (r \times l)/g$, where r corresponds to number of reads used in our analysis, l to read length and g to haploid genome size of *H. rhamnoides*.

Repeat Identification and Annotation

In order to identify repetitive sequences in the *H. rhamnoides* genome we employed comparative graph-based clustering analysis of sequenced reads by RepeatExplorer pipeline (Novak et al. 2013). Only clusters containing at least 0.01% of all clustered reads were considered and they corresponded to 58.5% of the genome. These were further manually characterized based on the similarity search results from RepeatMasker (http://www.repeatmasker.org; last accessed January 4, 2017) against Viridiplantae database and blastn and blastx (Altschul et al. 1990) against GenBank nr (Benson et al. 2009), which are part of the RepeatExplorer output. Cluster shapes were also used for repeat identification as tandem repeats with monomer longer than read length have typical donut-shaped clusters (Novak et al. 2010). Additionally, advanced analysis of satellite sequences, described in the section Satellite DNA sequences analysis, was used in the manual annotation of clusters.

Structural Annotation of LTR Retrotransposons

We reconstructed several Ty3/*Gypsy* and Ty1/*Copia* retrotransposons. The reconstruction comprised several steps. First, clusters belonging to particular element were visualized in SeqGrapheR (https://cran.r-project.org/web/packages/SeqGrapheR/index.html; last accessed January 4, 2017) program and contigs which together covered the whole elements were selected. These contigs were searched for occurrences of protein domains (GAG, RT, RH, AP, INT) by querying them to CDD (Marchler-Bauer et al. 2015). We then did multiple sequence alignment to create a consensus sequence of these contigs using progressive pairwise alignment implemented in Geneious 8.1.7 (http://www.geneious.com; last accessed January 4, 2017, Kearse et al. 2012). If necessary, resulting alignments were manually modified with respect to the order of domains for particular type of transposable element. The consensus sequence of reconstructed elements was then searched for the structural motif characteristics (ORFs and LTRs). Possible ORFs were detected by ORF Finder (https://www.ncbi.nlm.nih.gov/orffinder/; last accessed January 4, 2017). LTRs were determined on the basis of shape of a cluster and the element's coverage. Male and female coverage of reconstructed elements was determined by mapping reads which formed a current element to its consensus sequence using BowTie2 tool (Langmead and Salzberg 2012). Structural features and male and female coverage of reconstructed elements were visualized by custom R script and graph layouts of reconstructed elements were depicted by SeqGrapheR.

Phylogeny and Classification

Firstly, we created custom databases of plant LTR retrotransposon RT domains from sequences available in TREP (Wicker et al. 2002) and GyDB (Llorens et al. 2011) databases, independently for Ty3/*Gypsy* and Ty1/*Copia* retrotransposons. Contigs corresponding to retrotransposons were examined for the presence of a reverse transcriptase domain and Ty3/*Gypsy* and Ty1/*Copia* cores of RT domains were trimmed from these contigs based on the exact localization designated by CDD (Marchler-Bauer et al. 2015). Cores of RT domains were aligned by MUSCLE algorithm (Edgar 2004) together with our custom-made database of RT domains, and the resulting multiple sequence alignment was used as an input to create Neighbor-Joining tree (Saitou and Nei 1987) with Jukes-Cantor distance model using Geneious 8.1.7 (http://www.geneious.com; last accessed January 4, 2017, Kearse et al. 2012).

Preparation of Chromosomes and Probes and Fluorescence *In Situ* Hybridization

For chromosome preparations vegetatively propagated for commercial use, male ("Pollinator 1" and "Pollinator 3") and female (cv "Lomonosovskaya" and cv "Botanicheskaya ljubitelskaya") plants were used. Plant material was kindly provided by Dr G. Boyko, Lomonosov Moscow State University. The root tips were harvested separately from the individual male and female plants grown in pots. The harvested root tips were immediately pre-treated with a 2 mM aqueous solution of 8-hydroxyquinoline for 6 h at 20 °C. A 3:1 ethanol/glacial acetic acid (v/v) mix was used for fixation. Meristems 2 mm long were cut from the fixed root tips and digested in 10 µl enzyme solution [0.5% cellulase Onozuka R-10 (Serva, Germany) and 0.5% pectolyase Y-23 (Seishin Corp., Japan)] in 10 mM citrate buffer (pH = 4.9) for 2.5 h at 37 °C. The suspended cells were used for chromosome preparation as described by Kirov et al. (2014). The quality of spreads was assessed microscopically using phase-contrast and only preparations with at least 20 well-spread metaphases were used.

Probes for fluorescence *in situ* hybridization were generated using PCR-DIG Labeling Mix PLUS (Roche Diagnostics Gmbh) or by Biotin-11-dUTP 1/3 PCR labeling Mix (ZAO Sileks, Moscow). Primers for RT domain of selected transposable elements and determined monomer sequence of satellites were designed by Primer3 tool (Untergasser et al. 2012), were synthesized by ZAO "Syntol" (Moscow). These are available in supplementary table S1, Supplementary Material online. The pTa71 (45S rDNA) and pCT4.2 (5S rDNA) clones labeled by DIG-Nick translation kit were also used (Gerlach and Bedbrook 1979; Campell et al. 1992).

FISH experiments were performed as described in Alexandrov and Karlov (2016). For digoxigenin and biotin detection, slides were incubated with anti-DIG-FITC conjugate (Roche) and/or streptavidin-Cy3 conjugate (Sigma). The chromosomes were counterstained with DAPI (2 µg/ml) and mounted in Vectashield (Vector). An AxioImager M1 fluorescent microscope (Zeiss) was used to observe metaphase plates with fluorescent signals that were photographed with a monochrome AxioCam MRm CCD camera and visualized using Axiovision software (Zeiss).

Satellite DNA Sequences Analysis

As the seabuckthorn genome is abundant in satellite DNA and manual inspection would be exhaustive, we developed a custom bioinformatics approach which extended the basic analysis of RepeatExplorer tool. As an input the satellite clusters identified by RepeatExplorer are required. It is highly recommended to do manual inspection of these clusters and verify their structure and interaction with other clusters based on similarities among other clusters and pair-end reads connections. Our approach consisted of three basic steps.

(i) *Detection of satellite monomers*: First, assembled contigs of selected clusters were extracted from RepeatExplorer output and for each contig the monomer length was estimated from the distribution of distances between the same k-mers. The resulting monomer sequence was then extracted from the most covered part of the contig of previously determined length. Only the monomers with clearly distinguishable length, longer than 100 bp and reaching average coverage 50x and more were taken into account.

(ii) *Estimation of satellite families composition in genome and their annotation*: First, all to all monomer similarity was calculated. In order to do alignment of tandemly repeated monomers correctly (offsets between monomers are not known) we used one monomer as a subject and two copies in a row of the second monomer as a query. The similarity between monomers was then determined based on semiglobal alignment. To estimate the composition of satellite families in the genome, we clustered the monomer's similarity matrix using UPGMA method. The resulting dendrogram was then inspected by the user and cut off at the level that best discriminated the individual families (usually 70-85% of monomer identity). Identified families were visualized by the algorithm described by Fruchterman and Reingold (1991) implemented in igraph library and only connections that exceeded specified cut-off were considered and depicted. Secondly, to annotate identified families, all monomers were searched for similarity hits with sequences in the public nucleotide database and PlantSat database (Macas et al. 2002) using blastn (Altschul et al. 1990) with word size set to 11. Only results with an e-value lower than 10^{-20} were considered as significant. Finally, to depict satellite diversity inside the family, we chose the most covered monomer as a reference and mapped all reads belonging to the family onto its reference using BWA-MEM mapping tool (Li 2013). Conservation of different parts of the monomer was depicted using sequence logo created by WebLogo (Crooks et al. 2004) tool.

(iii) *Visualization of satellite families homogeneity:* First, the relative abundance of male and female reads was calculated in each tandem repeat family. This enabled us to predict their presence in sex chromosomes. We visualized the satellite homogeneity using the following procedure: reads from each identified family were merged together and sampled randomly to limit the maximum number of reads to speed up the following analysis. Similarity of sampled reads from all families was calculated using the megablast tool (Camacho et al. 2009) that performed all against all sequence comparison. Pairs of reads that met specific similarity threshold (70% sequence identity over at least 55% of sequence length) were further used for graph construction and visualization. Male and female reads were distinguished by color (male—blue, female—red), tandem repeat families were highlighted by different colors and the algorithm by Fruchterman and Reingold (1991) was used to depict the results. Additionally, graphs for selected families were refined with similarity thresholds ranging from 70% to 95% sequence identity to show satellite composition more clearly. Each satellite falling within individual satellite family was marked by a different color.

Results

Genomic Composition

We performed one Illumina MiSeq platform sequencing run for each male and female genomic DNA followed by graph-based clustering of reads and characterization of repetitive sequences by RepeatExplorer (Novak et al. 2013). All 223 clusters (with more than 167 reads) contained 973,049 reads corresponding to 58.5% of genome (fig. 1) and their identification showed that dominant (first) clusters corresponded to satellite DNA followed by Ty3/*Gypsy* and Ty1/*Copia* LTR retrotransposons. One cluster (CL97) corresponded to 5S rDNA, two clusters (CL40, CL71) to 45S rDNA and 15 clusters to chloroplast DNA (cpDNA). Although the majority of chloroplast DNA reads probably originated from contaminating cpDNA, some proportion could come from nuclear cpDNA insertions (NUPTs).

We identified main types of repetitive DNA and their genome proportions in male and female individuals (table 1). All transposable elements represented together 24% of male and 23% of female genome. Ty1/*Copia* retrotransposons formed 12%, Ty3/*Gypsy* retrotransposons 11% and DNA transposons 1.5% of male genome. The most abundant among Ty1/*Copia* retrotransposons were Angela/Tork and Ale/Retrofit, among Ty3/*Gypsy* retrotransposons Athila and chromoviruses dominated. No LINE elements were found in the whole seabuckthorn genome. Satellites together comprised about 27% of male and 24% of female genomes. The 45S rDNA formed 0.7% of both male and female genomes and 5S rDNA represented 0.2% of both male and female genomes.

Transposable Elements

To determine the phylogenetic relationships of Ty1/*Copia* and Ty3/*Gypsy* retrotransposons, we aligned their reverse transcriptase (RT) domains from individual clusters and constructed the phylogenetic trees. Both Ty3/*Gypsy* (fig. 2*A*) and Ty1/*Copia* (fig. 2*B*) trees contained families identified in our clusters (in red) mixed with representatives of known subfamilies of Ty1/*Copia* or Ty3/*Gypsy* from other plant species (in black). Among Ty3/*Gypsy* retrotransposons, we identified five clusters containing Athila subfamilies, one CRM subfamily, one Galadriel, one Reina and one Tat/Ogre subfamily (fig. 2*A*). Among Ty1/*Copia* retrotransposons, we found four subfamilies of Ale/Retrofit, four Angela/Tork subfamilies, one Maximus/SIRE subfamily, two TAR subfamilies and two Ivana/Oryco subfamilies (fig. 2*B*). The Angela/Tork and Ale/Retrofit subfamilies showed higher variability while Athila subfamilies were homogenous. Highest homogeneity were shown by chromoviruses where all reads were assembled into a single cluster for CRM, Galadriel and Reina families (fig. 2*A*).

We reconstructed the structure of the main Ty3/*Gypsy* and Ty1/*Copia* subfamilies (fig. 3) and identified all main features such as *gag* and *pol* genes (with all domains) and long terminal repeats (LTRs). In some retrotransposons (CL6, CL16) LTR regions were assembled into one long terminal repeat while in other clusters (CL7, CL27) right and left LTR were distinguished. This may be a consequence of lower or higher mutual diversity of LTRs in one element, and could correspond to age differences of elements. Graph layouts (right part of fig. 3) show the variability of specific parts of elements as well as alternative variants of elements, e.g., potential spliced variant (Novak et al. 2010). The similar coverage of elements by male and female reads indicates that elements are present on all chromosomes without accumulation/absence on

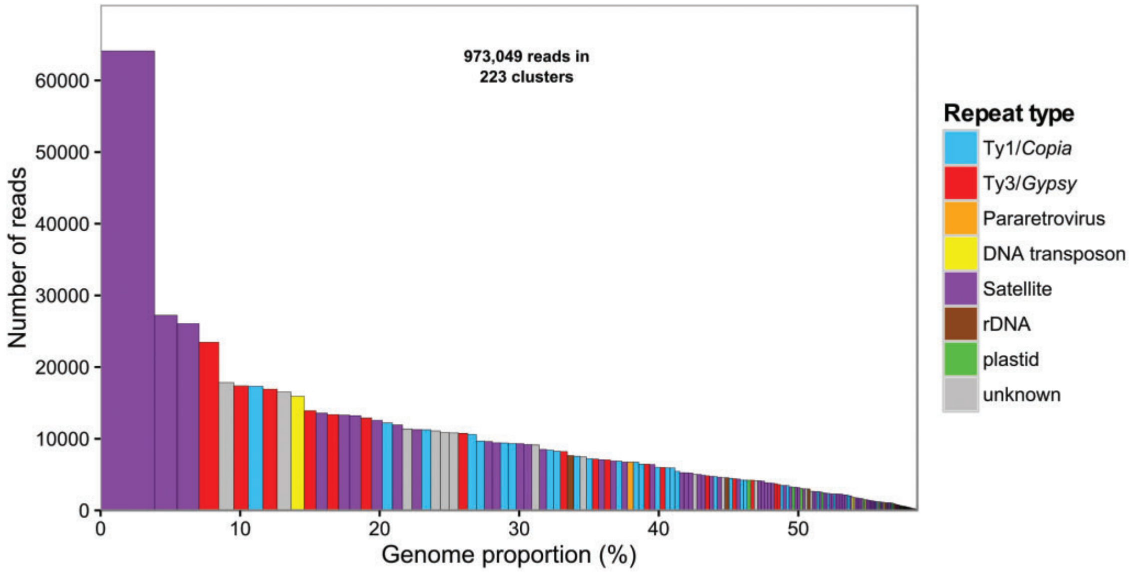


Fig. 1.—Repeat composition of clusters and their genomic proportions. Each column corresponds to one cluster and repeat types are distinguished by colors. The height of columns represents number of reads in each cluster, the width of column indicate genomic proportion of cluster.

Table 1
Repeat Composition in *Hippophae rhamnoides* Genome

Classification			Genome Proportion (%)	
Repeat Type	Super Family	Family	Male	Female
LTR retroelements	Ty1/Copia	Angela/Tork	4.83	4.90
		Ale/Retrofit	4.93	4.38
		TAR	1.34	1.06
		Maximus/SIRE	0.44	0.57
		Ivana/Oryco	0.25	0.23
		Total Ty1/*Copia*	11.79	11.15
	Ty3/Gypsy	Athila	6.39	5.36
		Chromovirus—CRM	2.98	3.58
		Chromovirus—Galadriel	1.28	0.80
		Chromovirus—others	0.27	0.31
		Chromovirus—Reina	0.06	0.04
		Tat/Ogre	0.05	0.05
		Total Ty3/*Gypsy*	11.04	10.15
DNA transposons			1.52	1.46
Total transposable elements			24.35	22.76
Pararetrovirus			0.48	0.59
rDNA	45S		0.77	0.69
	5S		0.20	0.16
Satellites			26.92	23.74
All repetitive elements			52.72	47.94
Unclassified			6.96	11.39
Low/single copy			38.96	39.50
Plastids			1.36	1.17

Note.—Types of repetitive DNA and their genome proportions.

the X or Y chromosome. Some elements had uninterrupted ORF corresponding to *gag* and *pol* (CL7, CL27, and CL43) and hence they can be active. Interruption of ORFs in other elements may have been caused by assembling errors during reconstruction (CL6, CL16, and CL37).

Satellite DNA

We developed a new bioinformatics approach for detailed analysis of satellite DNA in genomes. This method includes: (i) identification of satellite monomers based on distribution of distances of k-mers in assembled contigs, (ii) clustering of monomers allowing identification and annotation of satellite families in genome, and (iii) visualization of satellites homogeneity and male/female composition allowing better prediction of their localization with respect to sex chromosomes. Detailed description of the whole procedure is available in the section Materials and Methods and in supplementary figure S4, Supplementary Material online.

We utilized this approach for analysis of the seabuckthorn genome, but it is generally applicable in genomic studies of other species as well. As an input we used the 38 largest manually inspected satellite clusters from RepeatExplorer output extended by five smaller clusters with potentially interesting chromosomal localization (X, Y chromosomes). All clusters were grouped into 12 main superclusters that correspond to the 12 main families of satellite DNA in the seabuckthorn genome. Satellites were named HRTR1-HRTR12 (supplementary fig. S1, Supplementary Material online and table 2). Copy number of individual satellite families was determined based on following formula: $cn = [(s\ x\ l)/m]/cov$, where *s* represents number of reads of individual satellite family, *l* corresponds to

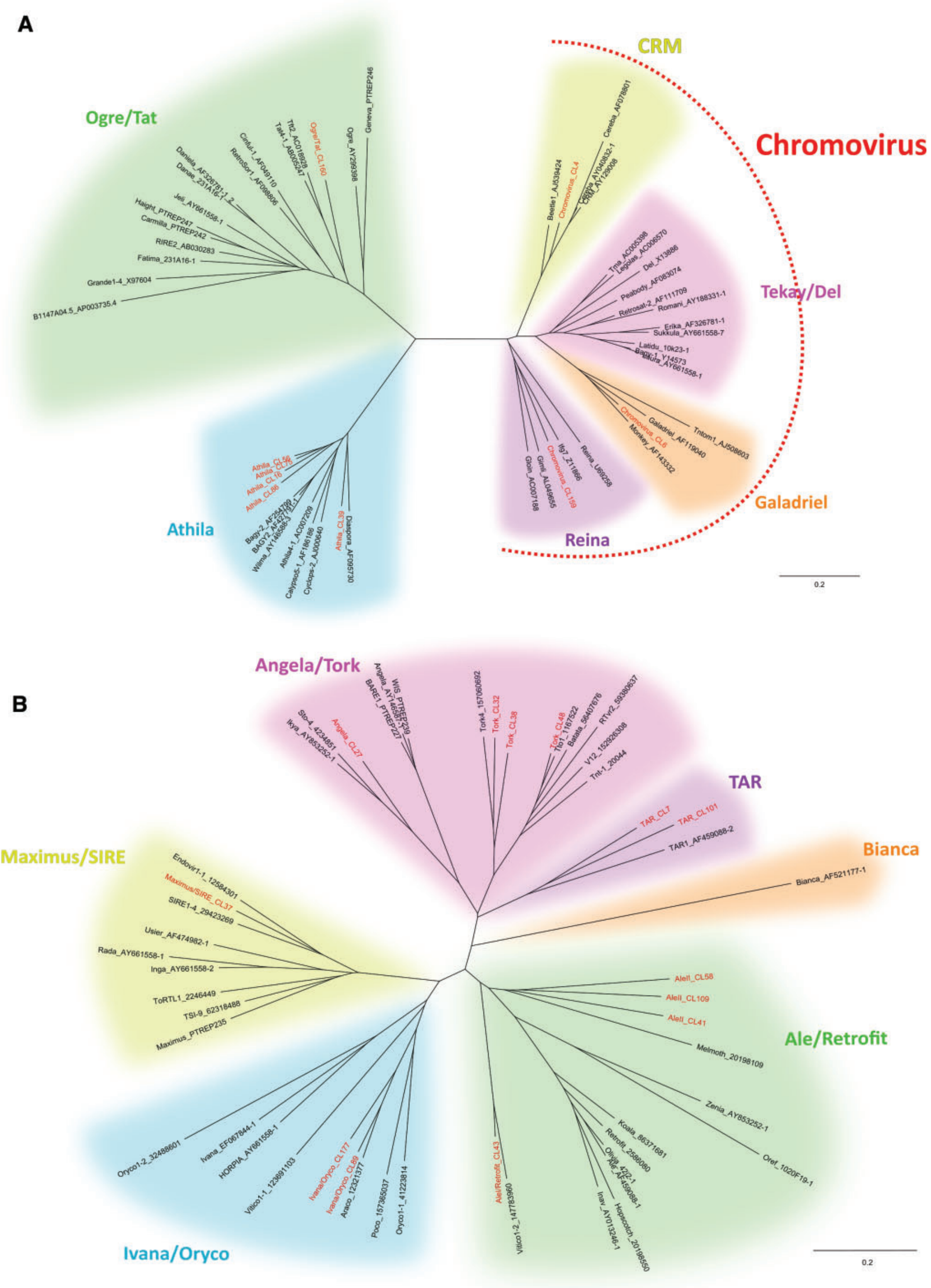


FIG. 2.—Phylogenetic trees of *Hippophae rhamnoides* Ty3/*Gypsy* (*A*) and Ty1/*Copia* (*B*) retrotransposons based on reverse transcriptase sequences. RT domains of retrotransposons reconstructed from Illumina reads in this study are in red, representative RT domains of retrotransposons from other plant species (from TREP and GyDB) are in black. Individual families are highlighted by different colors.

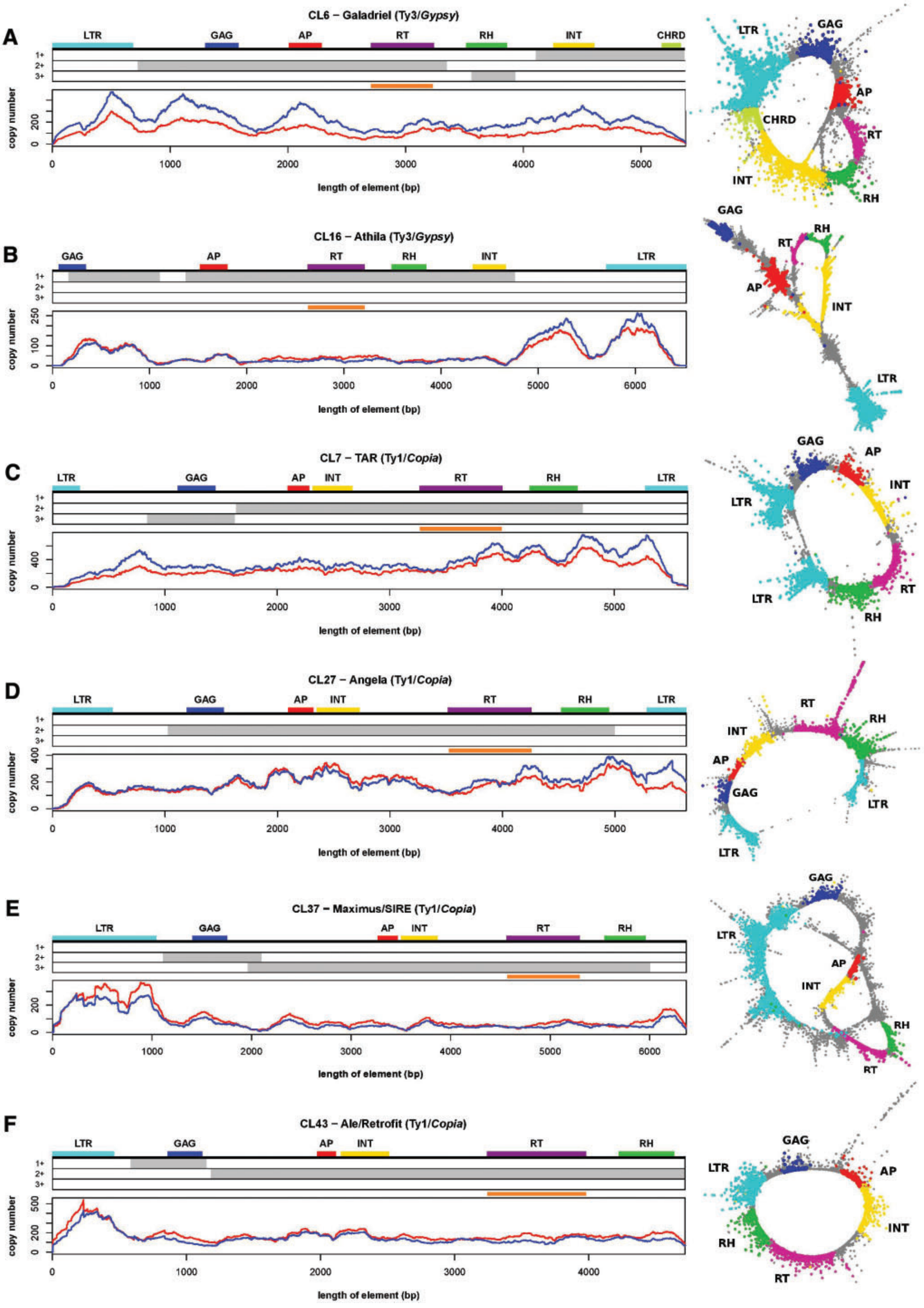


Fig. 3.—Comparison of structure of selected retrotransposon families in *Hippophae rhamnoides*. Graphs of coverage by male (in blue) and female (in red) genomic reads are showed under the structure of Ty3/*Gypsy* (*A*, *B*) and Ty1/*Copia* (*C*–*F*) elements shown in phylogenetic tree (fig. 2). Graph layouts on the right are visualized by SeqGrapheR program (http://cran.rproject.org/web/packages/SeqGrapheR/index.html). Protein domains and possible LTRs are distinguished by colors, found possible different three ORFs are marked by grey rectangles and orange line represents sequence for probes used for FISH.

Table 2
Main Satellite Families in *Hippophae rhamnoides* Genome

Name	Number of Reads	Localization	Monomer Length	M (%)	F (%)	Copy Number
HRTR1	129843	Strong signal on six pairs of small autosomes and weak signal on one pair of small autosomes	363	59.90	40.10	82270
HRTR2	60455	X and Y chromosome and weak signal on one pair of large and one pair of small autosomes	541	43.03	56.97	25702
HRTR3	46881	Dispersed signal on two large autosomal pairs	656	49.60	50.40	16437
HRTR4	27219	One pair of large and one pair of small autosomes	720	51.30	48.70	8695
HRTR5	23060	One pair of small autosomes	819	57.61	42.39	6476
HRTR6	19415	Three pairs of small autosomes	198[a]	53.67	46.33	5784[b]
HRTR7	14861	One pair of large autosomes and one pair of small autosomes	493[a]	68.38	31.62	4828[b]
HRTR8	12570	X chromosome and weak signal on one pair of small autosomes	826	35.06	64.94	3500
HRTR9	11155	One pair of small autosomes	354	69.52	30.48	7248
HRTR10	7476	Centromere of one pair of small autosomes	940	49.80	50.20	1829
HRTR11	4088	One pair of small autosomes	643	66.78	33.22	1462
HRTR12	1718	Y chromosome	257	100.00	0.00	1538

Note.—Names, monomer lengths, copy numbers, chromosomal localizations, and genome proportions.
[a]Shared length of the monomer in the family.
[b]Estimated based on average monomer length. 772 bp for HRTR6 and 708 bp for HRTR7.

read length, *m* represents estimated monomer length for satellite family and *cov* is genome coverage. Sequence logos show the monomer sequences of the main satellites and the sequence variability (supplementary fig. S2*A*–*L*, Supplementary Material online). Only HRTR1 and HRTR12 showed significant similarity hits with blast nucleotide (nr/nt) database (to previously deposited microsatellite markers of *H. rhamnoides*). There were no significant hits with PlantSat database for all satellite groups.

Based on our detailed analysis of HRTR6 and HRTR7, sharing small part of monomers (supplementary fig. S3*C*, Supplementary Material online), we decided to retain them as two separate tandem repeat families instead of one. These two families were very divergent and each showed variability in monomer' length (HRTR6: 730–810 bp, HRTR7: 475–830 bp). Monomers in each family had a common sequence (HRTR6: 198 bp, HRTR7: 493 bp) while other parts of monomers were significantly different from each other. For this reason, we only created sequence logos for the shared part of monomers for each family (supplementary fig. S2*F* and *G*, Supplementary Material online).

Male versus Female Comparison

To compare male and female genomes and to predict which repetitive DNA is specific for or accumulated on the X and Y chromosomes, we plotted the numbers of male versus female reads corresponding to individual clusters (fig. 4). This analysis involved all 223 clusters. The majority of clusters was located on the diagonal and these corresponded to transposable elements, rDNA and some satellites. However, some clusters containing satellites were enriched or even specific for males and represented potential Y-specific repeats. Other repeats, mostly satellites, were more abundant in females which could reflect their enrichment or specific localization on the X chromosome.

The greatest differences in composition of male and female reads were observed in satellites (five clusters located in the left; fig. 4). Detailed analysis showed that one of these (CL123—HRTR12) formed an isolated family composed of male reads only which suggests its localization only on the Y chromosome (fig. 5). The other four male biased satellites represented either a variant of a specific widespread cluster with Y chromosome presence (CL99 and CL144—HRTR2) or a satellite with a minor presence on the Y chromosome (CL150—HRTR1 and CL132—HRTR3). Eight satellites contained more female than male reads (2:1) indicating its localization on the X chromosome (female has two X chromosomes, male only one). HRTR2 satellite also contained more female than male reads but the ratio was 1.3 to 1 which could be explained by the localization on both sex chromosomes with greater abundance on the X than on the Y chromosome (fig. 5). Most other satellites had similar abundance of male and female reads, suggesting their localization (at least mostly) on autosomes.

Chromosomal Localization of Transposable Elements and Satellites

For determination of the chromosomal localization of transposable elements and satellites in seabuckthorn, we prepared probes representing reverse transcriptase region of individual TE families or part of a satellite monomer (supplementary fig.

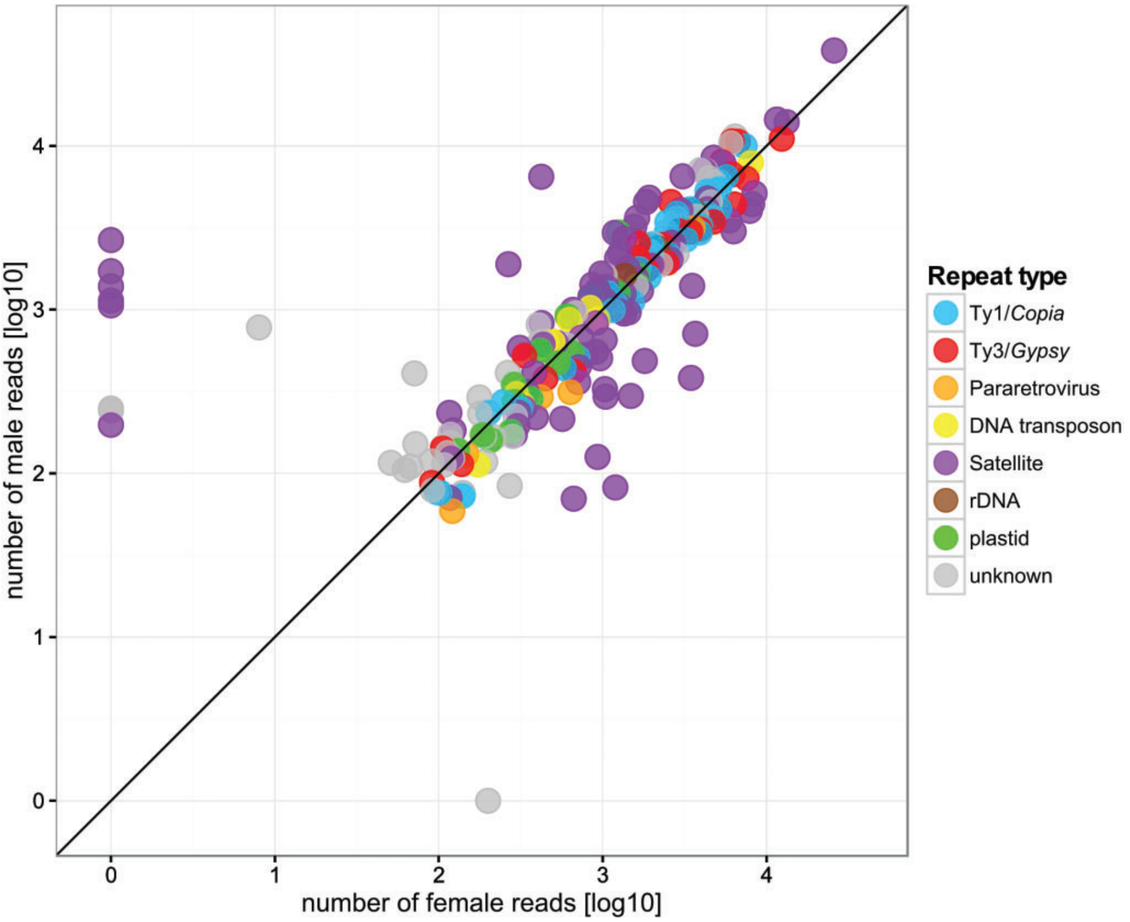


Fig. 4.—Comparison of repeats in male and female of *Hippophae rhamnoides*. Number of male versus female reads corresponding to individual clusters. Each circle in plot represents one cluster. Repeat types are marked by different color. Clusters in left upper part of graph are enriched (or specific) for males and thus potentially located on the Y chromosome while clusters in the right bottom part are enriched in female and thus potentially located on the X chromosome.

S1, Supplementary Material online) and used them for fluorescence *in situ* hybridization (FISH). In all FISH experiments we used both male (Pollinator 1, Leningradskaya region) and female (cv "Botanicheskaya lyubitelskaya") metaphases from plants that was used for sequencing. FISH experiments were also expanded to male ("Pollinator 3" Kaliningrad region) and female (cv "Lomonosovskaya"). In all ecotypes, we got the same results with X and Y.

FISH with satellite DNA showed various localization patterns on metaphase chromosomes of *H. rhamnoides* (fig. 6). The HRTR2, HRTR8 and HRTR12 show the sex specific or accumulation pattern of hybridization, while for HRTR3, HRTR4, HRTR5, HRTR6, HRTR7, HRTR9, HRTR10, and HRTR11 the hybridization patterns was the same for male as well as for female. The HRTR1 satellite hybridized mainly to heterochromatic arms of six pairs of small autosomes and weakly on one more pair of small autosomes (fig. 6*A* and *B*). In addition, a weak signal was detected distal to centromere on one arm of one large chromosome (chromosome X) in male (fig. 6*A*) and two large chromosomes in female (fig. 6*B*). The HRTR2 satellite gave a strong FISH signal on one large chromosome (chromosome X) and on one small chromosome (chromosome Y) in male (fig. 6*C*) and a strong FISH signal on two large chromosomes (chromosome X) in female (fig. 6*D*). Also a weak signal on the centromeric region of a pair of large and a pair of small autosomes was detected in both sexes. The HRTR3 satellite was localized on two large autosomal pairs with the FISH signal dispersed along these chromosomes (fig. 6*E*). The HRTR4 localized on one pair of large and on one pair of small autosomes (fig. 6*F*). The HRTR5 signal was detected on one pair of small autosomes only (fig. 6*G*). HRTR6 gave a strong signal on one autosomal pair and a weaker signals on two autosomal pairs (fig. 6*H*). The HRTR7 showed two sites of hybridization on one arm of a pair of large autosomes and on the centromeric region of a pair of small autosomes (fig. 6*I*). The HRTR8 hybridized mainly to the one large chromosome

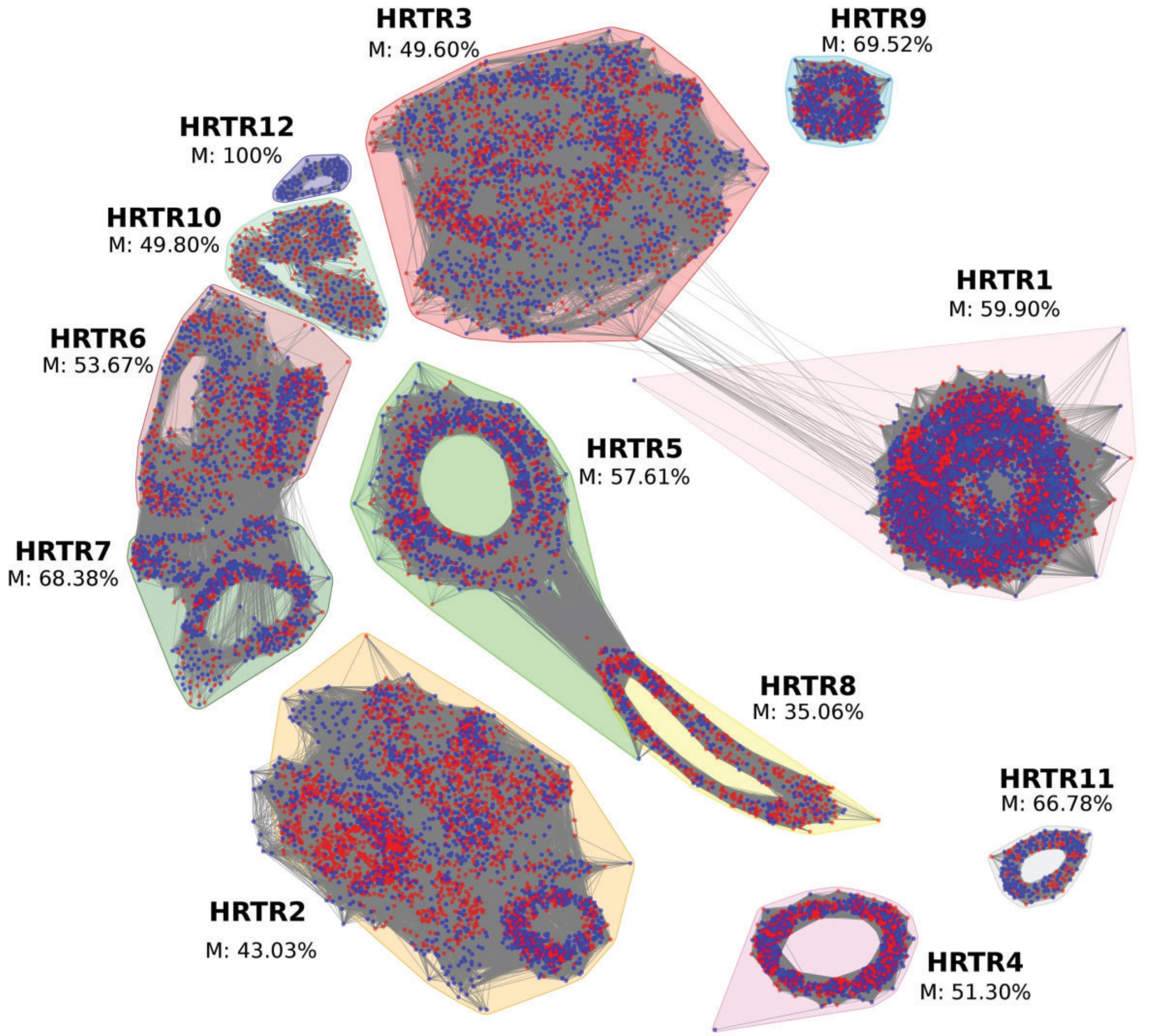


Fig. 5.—Visualization of male/female reads homogeneity in satellite families. Graph nodes correspond to sequenced reads and edges connect overlapping reads with more than 70% of sequence identity over at least 55% read length. Distances between reads are inversely proportional to their sequence similarity. Male reads are labeled by blue and female reads by red color. Individual families are highlighted by different colors. Please note HRTR12 family that is composed of male reads only assuming to be Y-specific.

(chromosome X) in male (Fig. 6*J*) and to the two large chromosomes (chromosomes X) in female (fig. 6*K*). A weak signal was also detected on one pair of small autosomes. The HRTR9, HRTR10, and HRTR11 were localized on one pair of small autosomes each (fig. 6*L–N*). The HRTR12 hybridized specifically to the small chromosome (Y chromosome) (fig. 6*O*) in male and no signal was detected in female (fig. 6*D*). The FISH signal intensity from HRTRs on X chromosomes varied depending on genotype.

Localization of the HRTR1 and the Y-specific (HRTR12), X-accumulated (HRTR8) and X and Y-accumulated (HRTR2) satellites on sex chromosomes was demonstrated by bicolor FISH using combinations of these probes and is summarized in a scheme (fig. 7). This together with specific or enriched representation of clusters in male and female (figs. 4 and 5), clearly demonstrates that *H. rhamnoides* has heteromorphic sex chromosomes (XY system) with large X and the small Y chromosomes.

We also mapped ribosomal genes. 45S rDNA was localized on one pair of small autosomes (fig. 8*A*) and 5S rDNA was localized on another pair of autosomes (fig. 8*B*). FISH with probes derived from transposable elements showed that

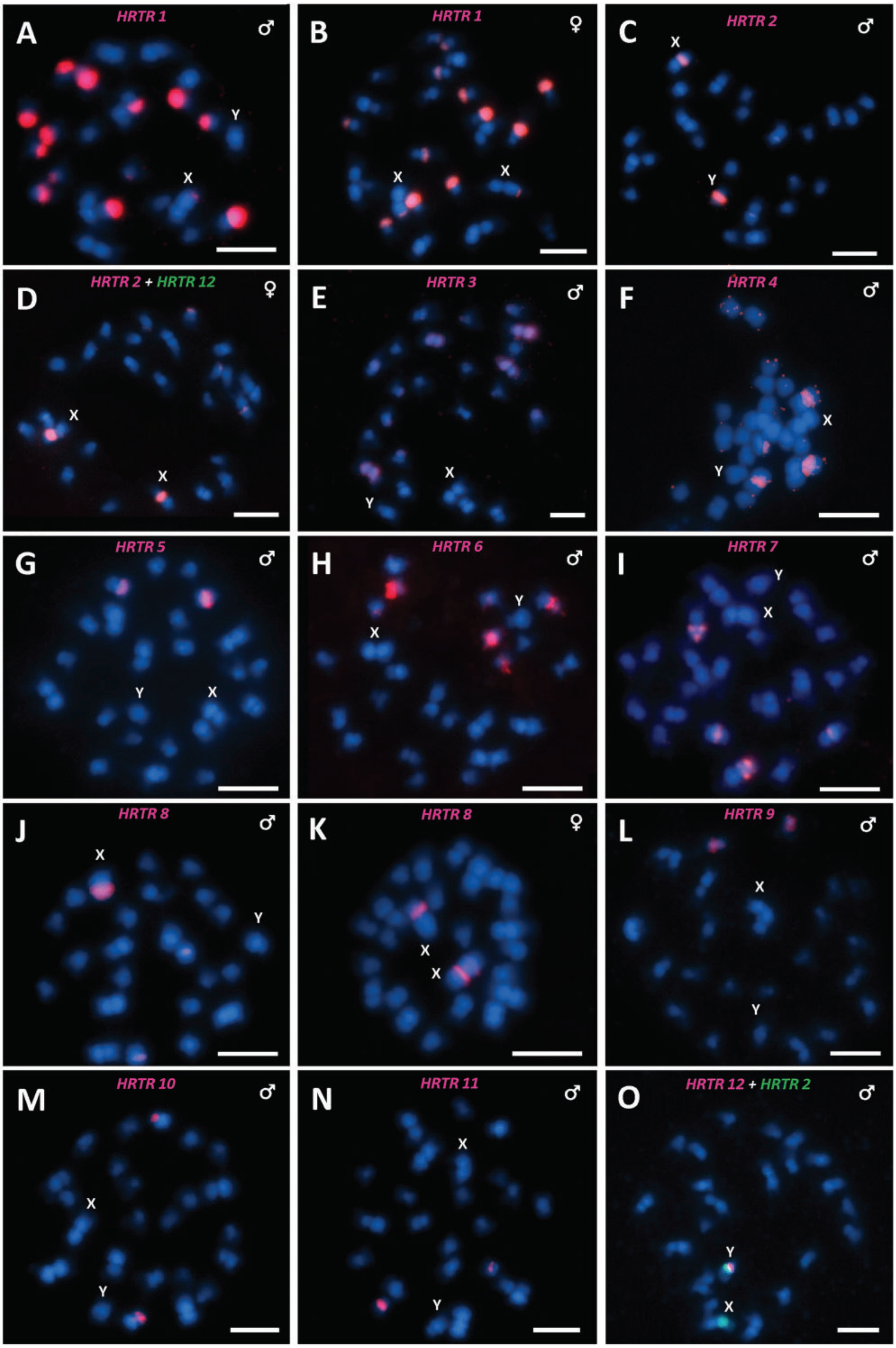


FIG. 6.—Localization of main satellite families on metaphase chromosomes of *Hippophae rhamnoides* using fluorescence *in situ* hybridization. The name of satellite family and sex of individual are indicated inside each figure. Blue are DAPI stained chromosomes, red and green signals show chromosomal localization of satellite families. Bar indicates 5 μm.

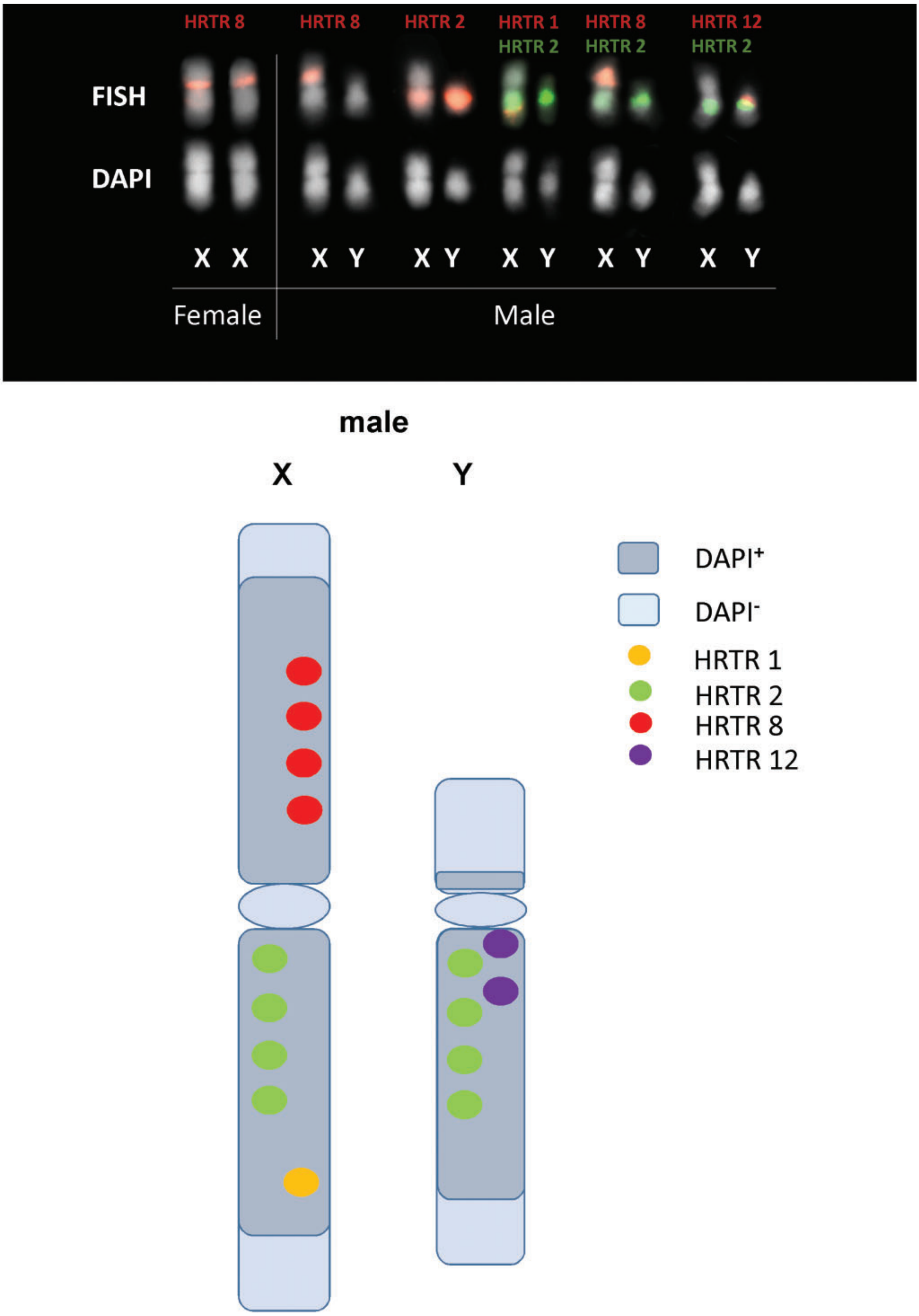


Fig. 7.—FISH and scheme of four satellites on sex chromosomes. The HRTR1, Y-specific HRTR12, X-accumulated HRTR8, sex chromosome-accumulated HRTR2.

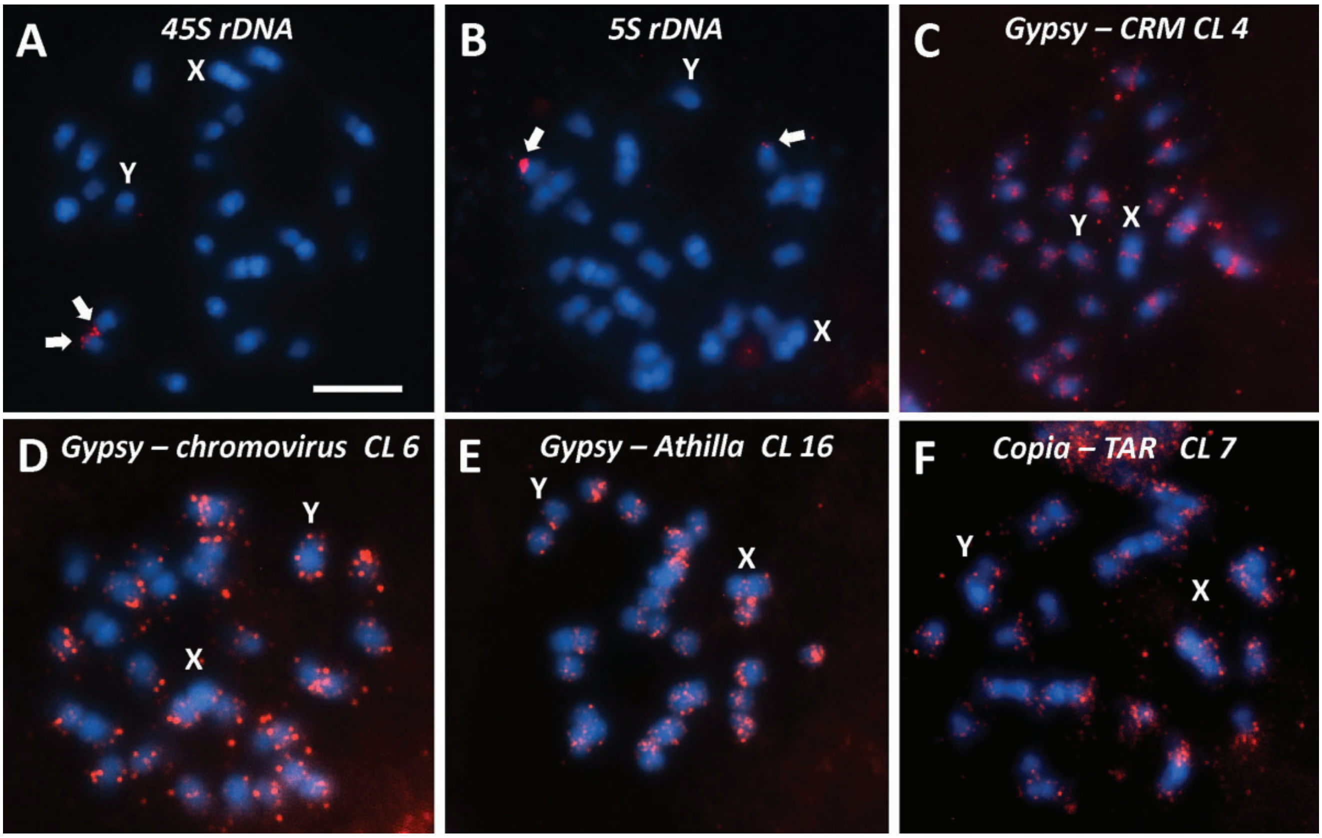


Fig. 8.—Localization of transposable elements and rDNA on metaphase chromosomes of *Hippophae rhamnoides* using fluorescence *in situ* hybridization. The name of transposable element family (together with the number of corresponding cluster) or type of rDNA cluster is inside each figure. Blue are DAPI stained chromosomes, red signal shows chromosomal localization of selected transposable elements and 45S and 5S rDNA. Bar indicates 5 µm.

three of four studied groups of TEs are present mainly in subtelomeres of all chromosomes (fig. 8*D–F*) and only the CRM retroelements (CL4) that was localized in the centromeric region of all chromosomes (fig. 8*C*).

Discussion

We present the first comprehensive analysis of seabuckthorn (*H. rhamnoides*) genome. We found that about one quarter of the genome is composed of TEs and another quarter of satellite DNA which is comparable to other plant genomes. Nevertheless, the seabuckthorn genome contains an unusually large number of different satellites (table 2, 12 main tandem repeats) compared with most other plant genomes (Mehrotra and Goyal 2014). Moreover, some satellites evolve rapidly into new variants. In particular, HRTR2 and HRTR3 satellite superclusters are comprised of a number of smaller clusters where each cluster represents an individual satellite (supplementary fig. S3, Supplementary Material online). Thus, the number of different satellites may be even higher if more strict criteria were used for tandem repeat classification. Transposable elements are represented by all main families of both Ty3/*Gypsy* and Ty1/*Copia* retrotransposons (fig. 2) with chromoviruses (CRM and Galadriel) and TAR families dominating (table 1). Most transposable element families are represented by only one or two clusters indicating their long term presence without changes in sequence or structure. Only Athila, Angela, Tork and Ale/Retrofit retrotransposons are found in multiple clusters (data not shown) suggesting higher divergence. Well preserved long ORFs in some TEs indicate the recent amplification/younger age and low level of degeneration of these elements. All in all, high variability of some satellites and TE families indicate high tempo of their diversification in the seabuckthorn genome, while other repeats remain relatively conserved. Nevertheless, this conclusion should be verified by comparative analysis of at least two closely related species. Recent analysis by Macas et al. (2015) showed that it is not transposable elements but satellites that are the most variable repeats among closely related species of Fabae genus.

Comparison of numbers of male and female reads constituting satellite superclusters, enabled us to predict satellites localized on the Y chromosome, X chromosome, on both sex chromosomes or on autosomes as each specific ratio of abundance of male and female reads in a cluster corresponded to specific chromosomal distribution. Our FISH results showed that this prediction works well in most cases as verified by satellites accumulated on the X chromosome (HRTR8) and both X and Y chromosomes, and specific for the Y chromosome (HRTR12) and for autosomes (HRTR1, 3, 4, 5, 6, and 10). It is a question whether or not the higher number of different satellites in the seabuckthorn genome than in the majority of plant genomes (Mehrotra and Goyal 2014) somehow correlates with the presence of sex chromosomes representing a specific genomic context, each shaped by different evolutionary forces.

The localization of satellites is remarkable and shows that satellites are gathered not only on the nonrecombining region of the Y chromosome but some are specific for the X chromosome or for both sex chromosomes. They are gathered in heterochromatic parts of sex chromosomes what can reflect possible role of satellites in heterochromatinization. The list of chromosomal localization of satellites and TEs in dioecious plants was recently presented by Li et al. (2016a). Although Y chromosome divergence and specific repeat composition is a generally accepted feature, an accumulation of X-specific repeats during plant sex chromosome evolution has been suggested only by limited number of studies (Hobza et al. 2004). As satellites localized on either X or Y chromosomes are mutually different, we prefer the explanation that these satellites originated and expanded on the sex chromosomes long after the X–Y divergence. Therefore, it would be interesting to compare X and Y-linked variants of HRTR2 satellite and, if present, to assess the extent of X- and Y-linked satellite divergence.

The localization of transposable elements mainly in subtelomeres is a feature characteristic of the seabuckthorn genome. However, transposable elements are accumulated in subtelomeres in other plant species too (Zhang and Wessler 2004), and, among dioecious plants, subtelomeric localization was shown in *Retand* retrotransposon in *Silene latifolia* (Kejnovsky et al. 2006). Retrotransposons are found in or around centromeres as well (Miller et al. 1998; Neumann et al. 2011).

Our results clearly confirm the existence of the XY system in seabuckthorn found by Shchapov (1979) and they show that the Y chromosome is small and the X chromosome large. We mention in passing the work of Truta et al. (2011) who initially found a large Y chromosomes and small X chromosome in three Romanian seabuckthorn genotypes that later investigation of Romanian genotypes failed to confirm (Dr. Elena Truta, Institute of Biological Research Iasi, Romania, personal communication, June 15, 2016). Another cytogenetic study on seabuckthorn using C-banding that unfortunately showed only female karyotype without marking sex chromosomes (Rousi and Arohonka 1980).

Estimation of the age of sex chromosomes is not yet possible in this species because no X- and Y-linked genes are known. It remains a question whether the large size difference between X and Y chromosomes, the small size of the Y chromosome and accumulation of different satellites on both sex chromosomes indicates greater age of these sex chromosomes or not. It is remarkable that another genus of the Elaeagnaceae family—Shepherdia (Elaeagnaceae contains three genera—Elaeagnus, Hippophae, and Shepherdia) contains only three species that are all dioecious (Veldkamp 1986). Moreover, the Elaeagnaceae family belongs to the order of Rosales containing other plants with heteromorphic sex chromosomes like Humulus and Cannabis. Although karyotypes were described in Elaeagnus ($2n = 28$ in *E. angustifolia*) and Shepherdia ($2n = 26$ in *S. argentea* and $2n = 22$ in *S. canadensis*), the sex chromosomes were not revealed (Rousi and Arohonka 1980). Therefore, it is not possible to draw conclusions about the formation or age of sex chromosomes during phylogeny.

The small Y chromosome containing several satellite DNA and a large X chromosome revealed in seabuckthorn resemble the mammalian sex chromosomal system. To the best of our knowledge, such a system is very rare among plants. Sex chromosomes in plants are mostly evolutionarily young—e.g., *Silene latifolia* (6 Ma, Kubat et al. 2014), *Rumex acetosa* (12–13 Ma, Navajas-Perez et al. 2005), or *Coccinia grandis* (3 Ma, Sousa et al. 2013)—and only sex chromosomes of *Marchantia polymorpha* are thought to be older (Yamato et al. 2007). A small Y chromosome and the large X chromosome were revealed in *Humulus lupulus* (Shephard et al. 2000; Karlov et al. 2003) and also in gymnosperm species *Cycas revoluta* (Segawa et al. 1971). The small size of the seabuckthorn Y chromosome may be caused by the loss of DNA which indicates that the Y chromosome could be in a shrinkage phase of evolution [reviewed in Hobza et al. (2015)] and thus could represent a rare example of an evolutionarily old plant sex chromosome. This assumption is supported by the FISH results which indicate that the large part of the Y chromosome arm that is homologous to the arm of the X chromosome, carrying HRTR8, was lost (fig. 7).

In this study, we developed and used a new bioinformatics approach for analysis of satellite DNA allowing prediction of satellite monomers, their grouping into clusters corresponding to main satellite families in the genome and visualization of their male/female homogeneity. This enabled prediction of satellite localization with respect to the sex determination system in species studied.

Supplementary Material

Supplementary data are available at *Genome Biology and Evolution* online.

Acknowledgments

This work was supported by the Czech Science Foundation (grant P501/12/G090). Access to computing and storage facilities owned by parties and projects contributing to the National Grid Infrastructure MetaCentrum, provided under the program "Projects of Large Infrastructure for Research, Development, and Innovations" (LM2010005), is greatly appreciated. The work of JP was supported by the Research and Application of Advanced Methods in ICT project (FIT-S-14-2299; http://www.fit.vutbr.cz/). The work of T.M. was supported from IT4Innovations excellence in science project (LQ1602). O.R., O.A., M.D., and G.K. were supported by Government Program supporting of the Leading Scientific Schools of Russian Federation (grant SS-8315.2016.11).

Literature Cited

Alexandrov OS, Divashuk MG, Yakovin NA, Karlov GI. 2012. Sex chromosome differentiation in *Humulus japonicus* Siebold & Zuccarini, 1846 (Cannabaceae) revealed by fluorescence in situ hybridization of subtelomeric repeat. Comp Cytogenet. 47:239–247.

Alexandrov OS, Karlov GI. 2016. Molecular cytogenetic analysis and genomic organization of major DNA repeats in castor bean (*Ricinus communis* L.). Mol Genet Genomics 291:775–787.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. J Mol Biol. 215:403–410.

Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. 2009. GenBank. Nucleic Acids Res. 38:D46–D51.

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30:2114–2120.

Camacho C, et al. 2009. BLAST+: architecture and applications. BMC Bioinformatics 10:421.

Campell BR, Song Y, Posch TE, Cullis CA, Town CD. 1992. Sequence and organization of 5S ribosomal RNA-encoding genes of Arabidopsis thaliana. Gene 112:225–228.

Cermak T, et al. 2008. Survey of repetitive sequences in *Silene latifolia* with respect to their distribution on sex chromosomes. Chromosome Res. 16:961–976.

Charlesworth B. 1991. The evolution of sex chromosomes. Science 251:1030–1033.

Charlesworth D. 2016. Plant sex chromosomes. Annu Rev Plant Biol. 67:397–420.

Chen L, Yu Z, Jin H. 2010. Comparison of ribosomal DNA ITS regions among *Hippophae rhamnoides* ssp. sinensis from different geographical area in China. Plant Mol Biol Rep. 28:635–645.

Crooks G, Hon G, Chandonia J, Brenner S. 2004. WebLogo: a sequence logo generator. Genome Res. 14:1188–1190.

Divashuk MG, Alexandrov OS, Kroupin PY, Karlov GI. 2011. Molecular cytogenetic mapping of *Humulus lupulus* sex chromosomes. Cytogenet Genome Res. 134:213–219.

Divashuk MG, Alexandrov OS, Razumova OV, Kirov IV, Karlov GI. 2014. Molecular cytogenetic characterization of the dioecious *Cannabis sativa* with an XY chromosome sex determination system. PLoS One 9:e85118.

Doležel J, Bartoš J, Voglmayr H, Greilhuber J. 2003. Nuclear DNA content and genome size of trout and human. Cytom Part A 51:127–128.

Doyle JJ, Doyle JL. 1990. Isolation of plant DNA from fresh tissue. Focus 12:13–15.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32:1792–1797.

Fruchterman TMJ, Reingold EM. 1991. Graph drawing by force-directed placement. Softw Pract Exp. 21:1129–1164.

Ghangal R, Chaudhary S, Jain M, Purty RS, Sharma PC. 2013. Optimization of de novo short read assembly of seabuckthorn (*Hippophae rhamnoides* L.) transcriptome. PLoS One 8:e72516.

Gerlach WL, Bedbrook JR. 1979. Cloning and characterization of ribosomal RNA genes from wheat and barley. Nucleic Acids Res. 7:1869–1885.

Hobza R, Lengerova M, Cernohorska H, Rubes J, Vyskot B. 2004. FAST-FISH with laser beam microdissected DOP-PCR probe distinguishes the sex chromosomes of *Silene latifolia*. Chromosome Res. 12:245–250.

Hobza R, Vyskot B. 2015. The genomics of plant sex chromosomes. Plant Sci. 236:126–135.

Hobza R, et al. 2015. Impact of repetitive DNA on sex chromosome evolution in plants. Chromosome Res. 23:561–570.

Hough J, Hollister JD, Wang W, Barrett SCH, Wright SI. 2014. Genetic degeneration of old and young Y chromosomes in the flowering plant *Rumex hastatulus*. Proc Natl Acad Sci U S A. 111:7713–7718.

Kearse M, et al. 2012. Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics 28:1647–1649.

Karlov GI, Danilova TV, Horlemann C, Weber G. 2003. Molecular cytogenetic in hop (*Humulus lupulus* L.) and identification of sex chromosomes by DAPI banding. Euphytica 132:185–190.

Kejnovsky E, et al. 2006. Retand: A novel family of gypsy-like retrotransposon harboring an amplified tandem repeat. Mol Genet Genomics 276:254–263.

Kejnovsky E, Hobza R, Kubat Z, Cermak T, Vyskot B. 2009. The role of repetitive DNA in structure and evolution of sex chromosomes in plants. Heredity 102:533–541.

Kejnovsky E, Vyskot B. 2010. *Silene latifolia*: the classical model to study heteromorphic sex chromosomes. Cytogenet Genome Res. 129:250–262.

Kirov I, Divashuk M, Van Laere K, Soloviev A, Khrustaleva L. 2014. An easy "SteamDrop" method for high quality plant chromosome preparation. Mol Cytogenet. 7:21.

Kubat Z, et al. 2014. Possible mechanisms responsible for absence of retrotransposon family on a plant Y chromosome. New Phytol. 202:662–678.

Langmead B, Salzberg S. 2012. Fast gapped-read alignment with Bowtie 2. Nat Methods 9:357–359.

Li H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv 1303.3997.

Li SF, Zhang GJ, Yuan JH, Deng CL, Gao WJ. 2016a. Repetitive sequences and epigenetic modification: inseparable partners play important roles in the evolution of plant sex chromosomes. Planta 243:1083–1095.

Li SF, et al. 2016b. DPTEdb, an integrative database of transposable elements in dioecious plants. Database 2016:1–10.

Lim KY, et al. 2007. Sequence of events leading to near-complete genome turnover in allopolyploid Nicotiana within five million years. New Phytol. 175:756–763.

Llorens C, et al. 2011. The Gypsy database (GyDB) of mobile genetic elements: release 2.0. Nucleic Acids Res. 39:D70–D74.

Macas J, Meszaros T, Nouzova M. 2002. PlantSat: a specialized database for plant satellite repeats. Bioinformatics 18:28–35.

Macas J, et al. 2015. In depth characterization of repetitive DNA in 23 plant genomes reveals sources of genome size variation in the legume tribe fabeae. PLoS One 10:e0143424.

Marchler-Bauer A, et al. 2015. CDD: NCBI's conserved domain database. Nucleic Acids Res. 43:222–226.

Mariotti B, Manzano S, Kejnovsky E, Vyskot B, Jamilena M. 2009. Accumulation of Y-specific satellite DNAs during the evolution of *Rumex acetosa* sex chromosomes. Mol Genet Genomics 281:249–259.

Mehrotra S, Goyal V. 2014. Repetitive sequences in plant nuclear DNA: types, distribution, evolution and function. Genomics Proteomics Bioinformatics 12:164–171.

Miller JT, Dong F, Jackson SA, Song J, Jiang J. 1998. Retrotransposon-related DNA sequences in the centromeres of grass chromosomes. Genetics 150:1615–1623.

Ming R, Bendahmane A, Renner SS. 2011. Sex chromosomes in land plants. Annu Rev Plant Biol. 62:485–514.

Navajas-Perez R, et al. 2005. The evolution of reproductive systems and sex-determining mechanisms within rumex (polygonaceae) inferred from nuclear and chloroplastidial sequence data. Mol Biol Evol. 22:1929–1939.

Neumann P, et al. 2011. Plant centromeric retrotransposons: a structural and cytogenetic perspective. Mobile DNA 2:4.

Novak P, Neumann P, Macas J. 2010. Graph-based clustering and characterization of repetitive sequences in next-generation sequencing data. BMC Bioinformatics 11:378–389.

Novak P, Neumann P, Pech J, Steinhaisl J, Macas J. 2013. RepeatExplorer: a Galaxy-based web server for genome-wide characterization of eukaryotic repetitive elements from next-generation sequence reads. Bioinformatics 29:792–793.

Rousi A, Arohonka T. 1980. C-band and ploidy level of *Hippophae rhamnoides*. Hereditas 92:327–330.

Saitou N, Nei M. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol. 4:406–425.

Segawa M, Kishi S, Tatuno S. 1971. Sex chromosomes of *Cycas revoluta*. Jpn J Genet. 46:33–39.

Sharma A, Zinta G, Rana S, Shirko P. 2010. Molecular identification of sex in *Hippophae rhamnoides* L. using isozyme and RAPD markers. For Stud China 12:62–66.

Shchapov NS. 1979. On the karyology of *Hippophaë rhamnoides* L. Tsitol Genet. 13:45–47.

Shephard HL, Parker JS, Darby P, Ainsworth CC. 2000. Sexual development and sex chromosomes in hop. New Phytol. 148:397–411.

Sousa A, Fuchs J, Renner SS. 2013. Molecular cytogenetics (FISH, GISH) of *Coccinia grandis*: a ca. 3 myr-old species of Cucurbitaceae with the largest Y/autosome divergence in flowering plants. Cytogenet Genome Res. 139:107–118.

Souza A, Bellot S, Fuchs J, Houben A, Renner SS. 2016. Analysis of transposable elements and organellar DNA in male and female genomes of a species with a huge Y-chromosome reveals distinct Y-centromeres. Plant J. 88:387–396.

Steflova P, et al. 2013. Contrasting patterns of transposable element and satellite distribution on sex chromosomes (XY_1Y_2) in the dioecious plant *Rumex acetosa*. Genome Biol Evol. 5:769–782.

Truta E, et al. 2011. Morphometric pattern of somatic chromosomes in three Romanian seabuckthorn genotypes. Caryologia 64:189–196.

Untergasser A, et al. 2012. Primer3–new capabilities and interfaces. Nucleic Acids Res. 40:1–12.

Veldkamp JF. 1986. Elaeagnaceae. In: Van Steenis CGGJ, de Wilde WJJO, editors . Flora Malesiana 10 (2), Martinus Nijhoff. Boston, London: The Hague, p. 151–156.

Wicker T, Matthews DE, Keller B. 2002. TREP: A database for Triticeae repetitive elements. Trends Plant Sci. 7:561–562.

Yamato KT, et al. 2007. Gene organization of the liverwort Y chromosome reveals distinct sex chromosome evolution in a haploid system. Proc Natl Acad Sci U S A. 104:6472–6477.

Zhang X, Wessler SR. 2004. Genome-wide comparative analysis of the transposable elements in the related species *Arabidopsis thaliana* and *Brassica oleracea*. Proc Natl Acad Sci U S A. 101:5589–5594.

Zhou X, et al. 2010. Genome size of the diploid hybrid species *Hippophae goniocarpa* and its parental species, *H. rhamnoides* ssp. sinensis and *H. neurocarpa* ssp. neurocarpa (Elaeagnaceae). Acta Biol Cracoviensia Ser Bot. 52:12–16.

Associate editor: Ellen Pritham

A.12 Paper XII

Guanine quadruplexes are formed by specific regions of human transposable elements

RESEARCH ARTICLE

Open Access

Guanine quadruplexes are formed by specific regions of human transposable elements

Matej Lexa[1†], Pavlina Steflova[2†], Tomas Martinek[3], Michaela Vorlickova[4,5], Boris Vyskot[2] and Eduard Kejnovsky[2*]

Abstract

Background: Transposable elements form a significant proportion of eukaryotic genomes. Recently, Lexa et al. (Nucleic Acids Res 42:968-978, 2014) reported that plant long terminal repeat (LTR) retrotransposons often contain potential quadruplex sequences (PQSs) in their LTRs and experimentally confirmed their ability to adopt four-stranded DNA conformations.

Results: Here, we searched for PQSs in human retrotransposons and found that PQSs are specifically localized in the 3'-UTR of LINE-1 elements, in LTRs of HERV elements and are strongly accumulated in specific regions of SVA elements. Circular dichroism spectroscopy confirmed that most PQSs had adopted monomolecular or bimolecular guanine quadruplex structures. Evolutionarily young SVA elements contained more PQSs than older elements and their propensity to form quadruplex DNA was higher. Full-length L1 elements contained more PQSs than truncated elements; the highest proportion of PQSs was found inside transpositionally active L1 elements (PA2 and HS families).

Conclusions: Conservation of quadruplexes at specific positions of transposable elements implies their importance in their life cycle. The increasing quadruplex presence in evolutionarily young LINE-1 and SVA families makes these elements important contributors toward present genome-wide quadruplex distribution.

Keywords: G4 quadruplex, Retrotransposons, Genome

Background

Transposable elements (TEs) are abundant inhabitants of eukaryotic genomes, representing e.g. about 50% of the human genome and up to 90% in some plant species. Long terminal repeat (LTR) retrotransposons are most common in plant genomes while animal genomes, including the human genome, are often flooded by non-LTR retrotransposons. Most of the human genome is transcribed and TEs therefore greatly contribute to cellular transcriptome and proteome [1,2]. Recent insertions of TEs underlie the variability of human populations and can cause several human diseases [3,4]. Somatic retrotranspositions occur during neuronal development [5,6] and tumorigenesis [7]. During the last two decades, it became widely accepted that TEs, as an inherently dynamic genome component, have an important role in both cell functioning [8] and genome evolution [9,10].

Human LTR retrotransposons are represented by endogenous retroviruses (HERV) but their activity is currently very limited: most HERVs were inserted into the genomes of our ancestors earlier that 25 mya [11]. LTR retrotransposons have LTR sequences at both ends, carry GAG and POL genes and several regulatory regions like promoter located inside LTR, primer binding site (PBS) and polypurine (PPT) sites where reverse transcription of the first and second strand of DNA starts, respectively. The majority of human TEs result from the present and past activity of non-LTR retrotransposons, including the LINE-1, Alu and SVA elements [8]. LINE-1 (long interspersed element 1, or L1) have two ORFs coding for RNA binding protein (ORF1) and endonuclease and reverse transcriptase (ORF2). ORFs are flanked with 5'-UTR and 3'-UTR regions. There are at least 850,000 L1 copies in the human genome [12]. Alu elements are about 300 bp long and have dimeric structure formed by the fusion of

*Correspondence: kejnovsk@ibp.cz
†Equal Contributors
[2]Department of Plant Developmental Genetics, Institute of Biophysics, Academy of Sciences of the Czech Republic, Královopolská 135, 61265 Brno, Czech Republic
Full list of author information is available at the end of the article

two monomers derived from 7SL RNA gene. Alus were active over the past 65 mya and the human genome contains more than 1 million copies. SVA elements are about 2 kb long and are composed of a hexamer repeat region, VNTR region, an Alu-like region, a HERV-K10-like region and polyadenylation signal ending with oligo(dA)-rich tail. SVAs were active throughout the last 25 mya of hominoid evolution and have about 3,000 copies [13]. Both Alu and SVA are trans-mobilized by the L1 machinery [14].

Molecular processes participating in the retrotransposon life cycle are regulated both by enzymes encoded by these elements themselves and by several host factors. It is probable that the activity of retrotransposons can also be affected by the changes of DNA conformation that are known to influence many molecular processes (for review see [15]). Formation of multi-stranded DNA structures, namely quadruplex DNA, is probably involved in dimerization of the HIV-1 genomic RNA molecules found in virus particles [16]. Similarly, long polypurine tract (PPT) located in 3'-UTR of L1 retrotransposons, where reverse transcription of the second cDNA strand starts, can form intrastrand quadruplex [17]. Relationship between quadruplexes and transposons can be seen in the cleavage of quadruplexes by RAG1 protein during translocations in human lymphomas [18] because RAG1 protein evolved from transposase of the Transib family of DNA transposons [19].

Recently, we found [20] that potential quadruplex sequences (PQSs) are often located inside LTRs of plant LTR retrotransposons at specific distances from their promoter indicating a possible effect of quadruplexes on transcription. Quadruplexes were better preserved in evolutionary young elements which supports their functional role [20,21]. Similar observation was made by Savage et al. [22] who found that younger human SVA elements contain more PQS sequences than older SVA elements but the ability of candidate sequences to adopt quadruplex conformation was not experimentally confirmed. Although quadruplexes were found in many regions of human genome, especially in promoters [23-25], systematic analysis of quadruplexes in all main types of human retrotransposons was lacking.

In this study, we searched for PQS sequences in human LINE-1, HERV, SVA and Alu elements. We analyzed the prominent regions of their location as well as the effect of element age and localization on chromosomes. The ability of candidate motifs to adopt quadruplex was verified by circular dichroism and gel electrophoresis.

Results

Potential quadruplex-forming sequences are located in specific regions of human transposable elements

We analyzed the localization of PQSs inside main groups of human transposable elements (TEs), namely in LINE-1, Alu elements, HERV retrotransposons and SVA elements. We searched for the $(G_nX_nG_nX_nG_nX_nG_n)$ motif representing potential G-quadruplex cluster inside 894,717 LINE-1 elements, 1,051,161 Alu elements, 38,578 HERV and 5,001 SVA elements or their fragments. Altogether, we found 264,711 PQS in all annotated repeats or their 200 bp flanking sequence (186,507 in plus strands, 78,204 in minus strands). Of those, 183,967 were associated with the four studied classes (136,977 in plus strands, 46,990 in minus strands).

The overall highest abundance of PQSs was observed in SVA (PQS was in 36.2% of elements) followed by LINE-1 (PQS in 7.7% of elements) and HERV elements (PQS in 4.8% of elements). The occurrence of PQSs was lowest in Alu elements (in 1.1% of elements) (a complete list of TE families that contribute more than 1% of PQSs present in the entire genome is available as Additional file 1), showing PQS distribution and frequency in RepeatMasker subfamilies of Alu, ERVL-MaLR, ERVL, ERV1, haT-Charlie, L1, L2, MIR and SVA elements). In LINE-1 elements, PQSs were located almost exclusively in the 3'-UTR region. Only very low numbers of PQSs were found outside this region (Figure 1a). HERV LTR retrotransposons contained PQSs along the whole element with accumulation in LTR regions (Figure 1b). In SVA elements, PQSs were specifically located in Hex region in minus strand and along the larger VNTR region in plus strand (Figure 1c). The occurrence of PQSs inside Alu elements was low throughout most of element length (Figure 1d). There was only one peak of PQS in the left part of left monomer (50 bp from the 5'-end). All mentioned PQS peaks were above the Markov model random background threshold, with SVA-VNTR region being much closer to it than the other PQSs, as expected for a long G-rich tandem repeat.

We analyzed the abundance of PQSs in LINE-1, HERV, SVA and Alu elements separately on the Y and X chromosomes and on autosomes. In four main types of elements, the number of quadruplexes was measured on both plus and minus strands for respective chromosomes. The distribution of PQSs along all elements slightly differed between chromosomes being more similar in autosomes and X chromosome and different in the Y chromosome where peaks of PQSs were located in different parts of elements than in autosomes and the X chromosome (Figure 1). The most striking difference between PQS distribution or frequency was observed in the SVA family of transposable elements, where those on the Y chromosome had a reduced PQS content (Figures 1c and 2b). Intriguingly, we noticed an increased occurrence of PQSs in the ORF2 region of LINE-1 elements from chromosome X and Y compared to their autosome counterparts (Figure 1a).

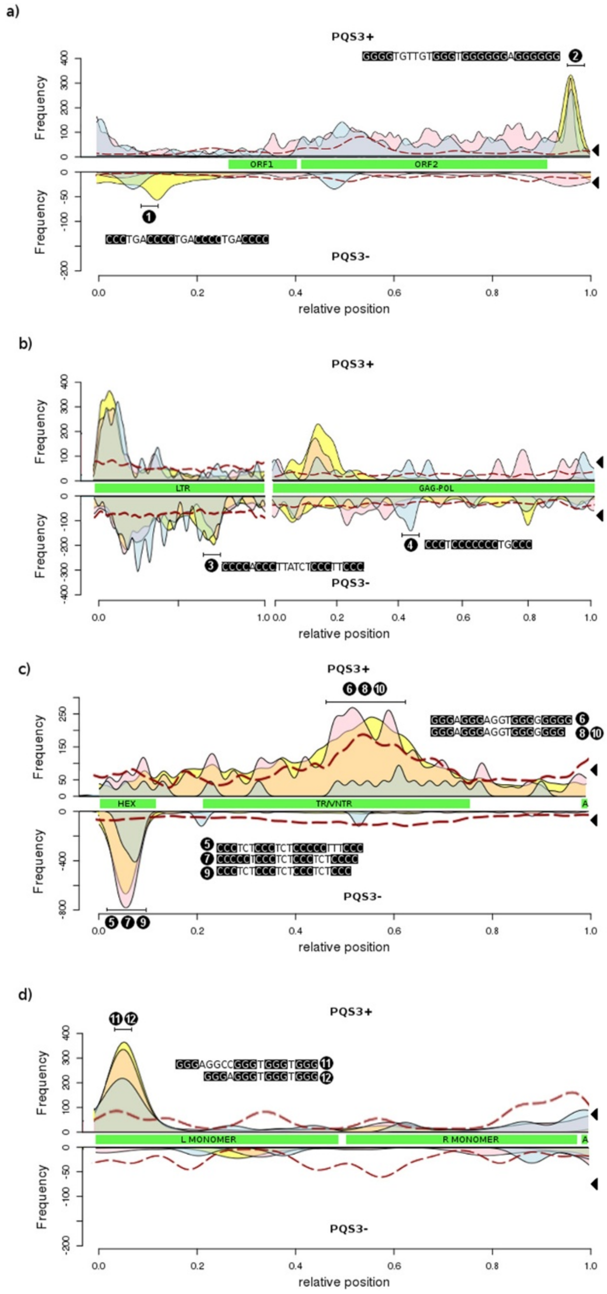


Figure 1 Occurrence of PQSs along human LINE-1 (a), HERV (b), SVA (c) and Alu (d). The density of PQS clusters containing a minimum of four adequately spaced GGG groups in the sense strand (PQS3+, upper lines) and antisense strand (PQS3-, lower lines) visualized along LINE-1 **(a)**, HERV **(b)**, SVA **(c)** and Alu elements **(d)**. Sliding window covered between 40-120 bp of the element length. Frequency represents the number of PQSs in such window in the entire family. Green boxes show annotation with main structural components from typical full-length elements (ORF - open reading frame, LTR - long terminal repeat, Hex - hexamer tandem repeat with a CCCTCT consensus, TR/VNTR - SVA tandem repeats with a period of approximately 36, A - polyA tail, L and R MONOMER - 7SL RNA-derived monomer). Chromosomes are visualized separately where autosomes are in yellow, X chromosomes in red and Y chromosomes in blue. The dashed red line shows PQS frequency in randomized control sequences generated by an equivalent 2nd-order Markov chain model. The black triangles on the right show reference densities of PQS sites in the entire human genome recalculated into the coordinates of the given family.

We clustered PQSs from individual families to determine the most common patterns of guanines. We have chosen 2 specific motifs for each TE type among the most common PQS motifs and used them for DNA conformational studies (Table 1). The selected sequences originated from the 5'-UTR and 3'-UTR regions of

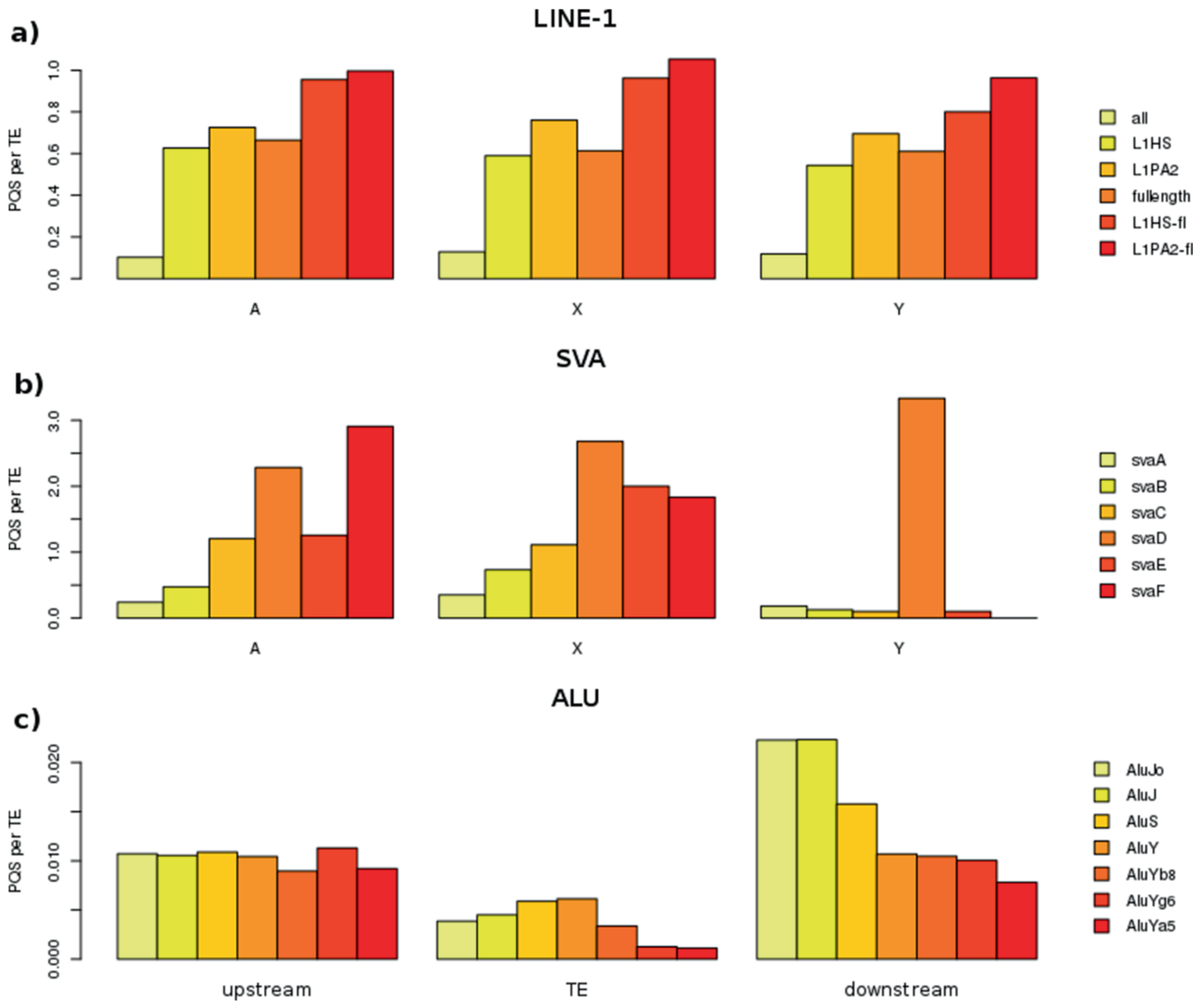


Figure 2 The abundance of PQSs inside LINE-1, SVA and Alu elements related to the element length, activity or age. (a) The abundance of PQSs inside all LINE-1, full-length LINEs and active LINEs (L1HS and L1PA2) located on autosomes and X and Y chromosomes. **(b)** The abundance of PQSs inside SVA of different age (oldest SVA-A to youngest SVA-F) located on autosomes and the X and Y chromosomes. **(c)** The abundance of PQSs upstream inside and downstream of Alu elements of different age (oldest Alu-J, middle-aged Alu-S and youngest Alu-Y) located on autosomes and the X and Y chromosomes.

LINE-1, LTR and gag-pol gene of HERV, Hex and VNTR regions of SVA and left part of the left monomer of Alu (Figure 1a-d).

The abundance of PQSs in the neighborhood of transposable elements

We compared the abundance of PQSs inside and in the vicinity of LINE-1, HERV, SVA and Alu elements. In full-length LINE-1 elements, the density of PQSs was markedly higher inside elements than in element vicinity. The greater abundance of PQSs inside LINE-1 elements compared to the element vicinity was observed only in plus strand while the neighborhood contained more PQSs than element when minus strand was analyzed (Figure 3a). Elements with the 3'-UTR PQS were much less likely to have the PQSs in the 3' downstream flanking region (data not shown). In HERV elements, many more PQSs were present inside elements than in their neighborhood, especially when full-length elements were taken into account (Figure 3b). High enrichment of elements compared to their neighborhood was also observed in SVA elements (Figure 3c). This trend was stronger in plus than in minus strand. In minus strand, SVA contained more PQSs upstream than downstream of elements. Alu elements differed from LINE-1, HERV and SVA. The density of PQSs inside Alu elements was lower than in regions located upstream and downstream (Figure 3d).

Table 1 Oligonucleotides used in this study

Number	Name	Sequence	Length [nt]
1	L1_1	TAGGTGCTC **GGGG** TCA **GGGG** TCA **GGGG** TCA **GGG** ACCCACTTG	42
2	L1_2	ATCACACTCT **GGGG** TGTTGT **GGG** T **GGGGGG** A **GGGGGG** AGGATAGCATT **GGG** AGATATACC	60
3	HERV_1	AAAGAGTCA **GGG** AA **GGG** AGATAA **GGG** T **GGGG** CCGTTTTAT	40
4	HERV_2	TAAATTGCT **GGG** CA **GGGGGGG** A **GGG** CTAGTCACG	34
5	SVA-A_HEX	GGAGATCAA **GGG** AAA **GGGGG** AGA **GGG** AGA **GGG** AGAGGCCAA	41
6	SVA-CF_VNTR	CGCCCGTCC **GGG** A **GGG** AGGT **GGGGGGGG** TCAGCCCCC	37
7	SVA-C_HEX	GGAGACCGT **GGGG** AGA **GGG** AGA **GGG** A **GGGGG** AGAGGAGAC	40
8	SVA-BF_VNTR	GCCCCGTCC **GGG** A **GGG** AGGT **GGGGGGG** TCAGCCCCC	36
9	SVA-F_HEX	GGAGAGAGA **GGG** AGA **GGG** AGA **GGG** AGA **GGG** AGA **GGG** AGAGTGCTG	45
10	SVA-F_VNTR	GTGCCATCC **GGG** A **GGG** AGGT **GGGGGGG** TCAGCCCCC	36
11	ALU-S_1	CCAGCACTTT **GGG** AGGCC **GGG** T **GGG** T **GGG** TCACCTGAGG	39
12	ALU-S_2	CCAGCACTTT **GGG** A **GGG** T **GGG** T **GGG** TGGATCACTT	35

Names and the sequences of oligonucleotides are shown. Clusters of three or more guanines are shown as bold.

The abundance of PQSs within transposable elements of different age and activity

We compared the PQS abundance in all LINE-1 elements, full-length LINE-1 and transcriptionally active LINE-1 families (L1HS and L1PA) [26]. We found that full-length LINEs contained much more PQSs than truncated LINE elements (Figure 2a). Among full-length elements, the transcriptionally active L1HS and L1PA2 families contained more PQSs than was the average abundance of PQSs inside full-length LINEs. Truncated L1HS and L1PA2 homologues contained much less PQSs. These trends were observed both on autosomes and on X and Y sex chromosomes.

We analyzed the abundance of PQSs inside SVA elements of different age - SVA-A (oldest family) to SVA-F (youngest family). We found that the abundance of PQSs was higher in younger elements (SVA-D, SVA-E and SVA-F) than in older elements (SVA-A, SVA-B and SVA-C) and this trend was same both in autosomes and sex chromosomes (Figure 2b). The abundance of PQSs was highest in middle-aged SVA elements (Figure 2b). The PQSs were common in the central part of elements in plus strand. Detailed analysis revealed that in older elements, the PQS abundance in the central part of plus strand decreased and predominated in the left part of SVA in the minus strand (not shown). The peak of PQSs in SVA-E present on the Y chromosome was caused by the low number of elements and the SVA-F elements even absented on the Y chromosome.

We made similar analysis of Alu elements where Alu-J are oldest, Alu-S are middle-aged and Alu-Y are youngest elements. We found that in contrast to LINE-1 and SVA, the age did not markedly affect the abundance of PQSs inside Alu elements. There is a slight PQS-increasing trend with age in the main families, however the youngest subfamilies (AluYg6, Ya5) [27] are also depleted of PQSs (Figure 2c). Because Alu elements have more PQSs in their vicinity than inside elements (Figure 3) we also analyzed the upstream and downstream regions. We found that older Alu elements contained more PQSs than younger elements in their downstream regions (Figure 2c). The PQS abundance did not differ markedly between autosomes and sex chromosomes, a small decrease in PQSs on the Y chromosome was registered (Figure 1d). The most active families of Alu (AluYg6 and AluYa5) had lower abundance of PQSs than average Alu-Y elements.

PQSs can form quadruplexes as revealed by circular dichroism

We probed DNA conformational properties of 12 oligonucleotides (Table 1) representing PQSs obtained from SVA, HERV, LINE-1 and Alu elements by circular dichroism (CD). We tested their ability to form quadruplex structures upon increasing concentration of potassium ions.

First, we measured CD spectra of PQSs originating from SVA elements because they contain PQSs more often than any other human TEs. We divided SVA elements into three families with different age - oldest SVA-A family, middle-aged SVA-C family and youngest SVA-F family - and for each family we analyzed the ability of one Hex region and one VNTR region to adopt a quadruplex structure. VNTR consensus sequences in older families were always present in younger families as well, therefore we used consensus sequences common for multiple families - SVA-BF for families B to F and SVA-CF for families C to F. Only SVA-F VNTR oligonucleotide was specific for the youngest SVA family. As shown in Figure 4a, the positive CD band at about 260 nm, which is characteristic of the presence of a parallel quadruplex [28], increased

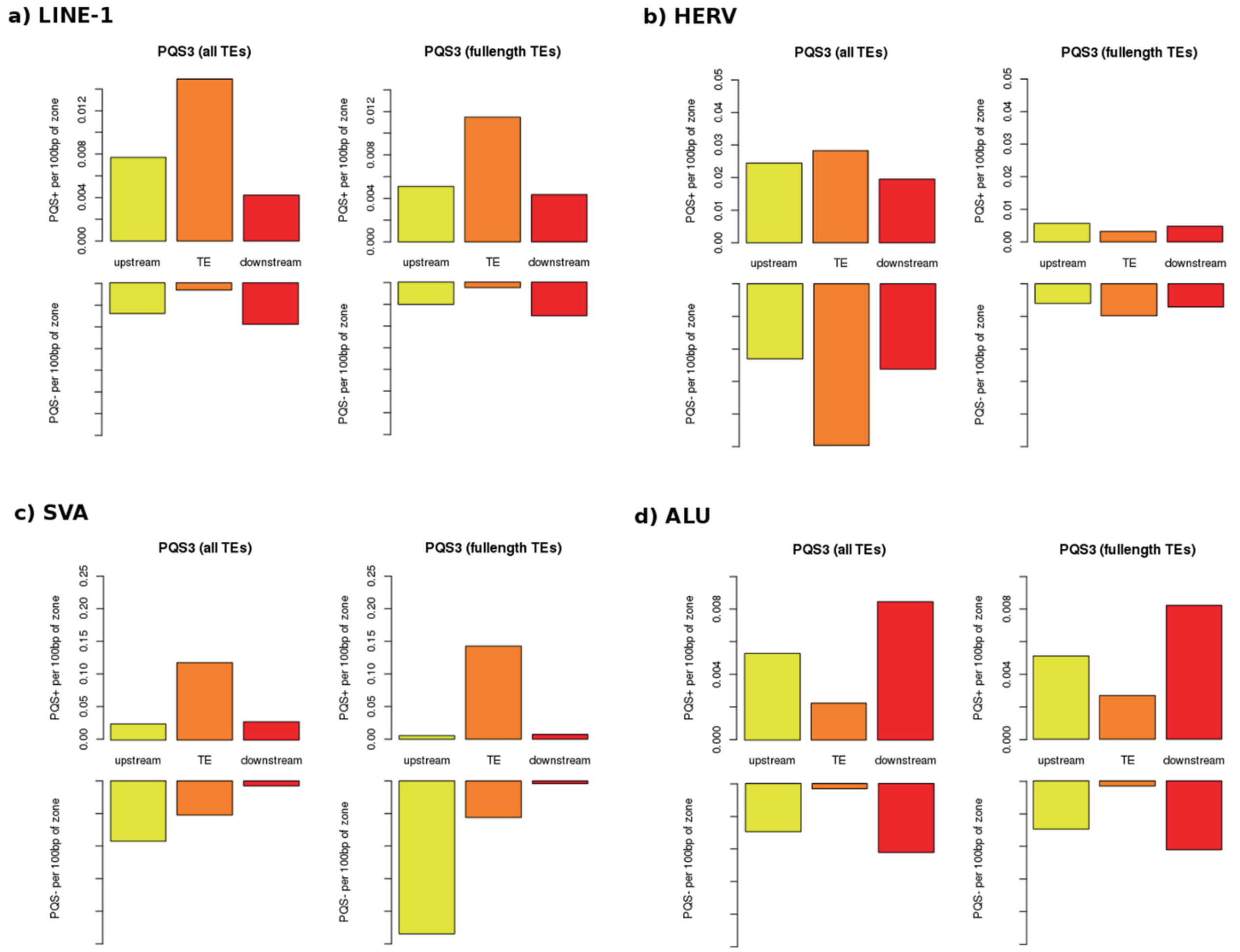


Figure 3 The density of PQSs inside TEs and in TEs vicinity. The number of PQSs per 100 bp of zone located inside, upstream and downstream of LINE-1 **(a)**, HERV **(b)**, SVA **(c)** and Alu elements **(d)**. For each element type PQS in plus strand (upper row) and minus strand (lower row) are shown. We analyzed either all elements (left figures) or only full-length elements (right figures).

steeply and at lower potassium concentrations with SVA-F (youngest) than with SVA-C (Figure 4a). Much less increase in this band and only at the highest K^+ concentrations used was observed with the older SVA-A family and the common consensus oligonucleotides SVA-BF and SVA-CF. Similarly, the thermal stability of quadruplexes was highest in SVA-F and lowest in SVA-A (not shown). Native gel electrophoresis at 150 mM K^+ showed that Hex region of three groups of SVA adopted bimolecular quadruplexes (Figure 4b). VNTR region provided CD spectra of the B-DNA type at low K^+ concentrations marked out by low amplitudes and a slightly predominating 260 nm band, which is characteristic of duplexes of G-rich and C-rich DNA strands [29]. These monomolecular structures (Figure 4b) may thus correspond to hairpins containing rather accidental, namely G.C, base pairs. The increase in the 260 nm band with increasing K^+ concentration indicating quadruplex formation was again most obvious with SVA-F and less so with SVA-BF and SVA-CF. The quadruplexes were formed non-cooperatively and not much willingly.

PQSs originating from the LINE-1 elements were selected from the 5'-UTR and 3'-UTR regions (Figure 1). The PQSs from the 5'-UTR (labelled as L1_1) provided CD spectrum corresponding to antiparallel quadruplex (Figure 5a), while the CD spectrum of the PQS from 3'-UTR (labelled as L1_2) corresponded to that of the parallel-stranded quadruplex (Figure 5a). Native PAGE revealed that both the anti-parallel L1_1 and the parallel L1_2 quadruplexes were monomolecular at low as well as at room temperature (Figure 5b). Antiparallel folding of the L1_1 quadruplex was enabled by the sufficiently long (trinucleotide) loops between all four G blocks.

Two PQSs were selected from HERV elements. The first PQS corresponded to a minor PQS peak in the LTR in the

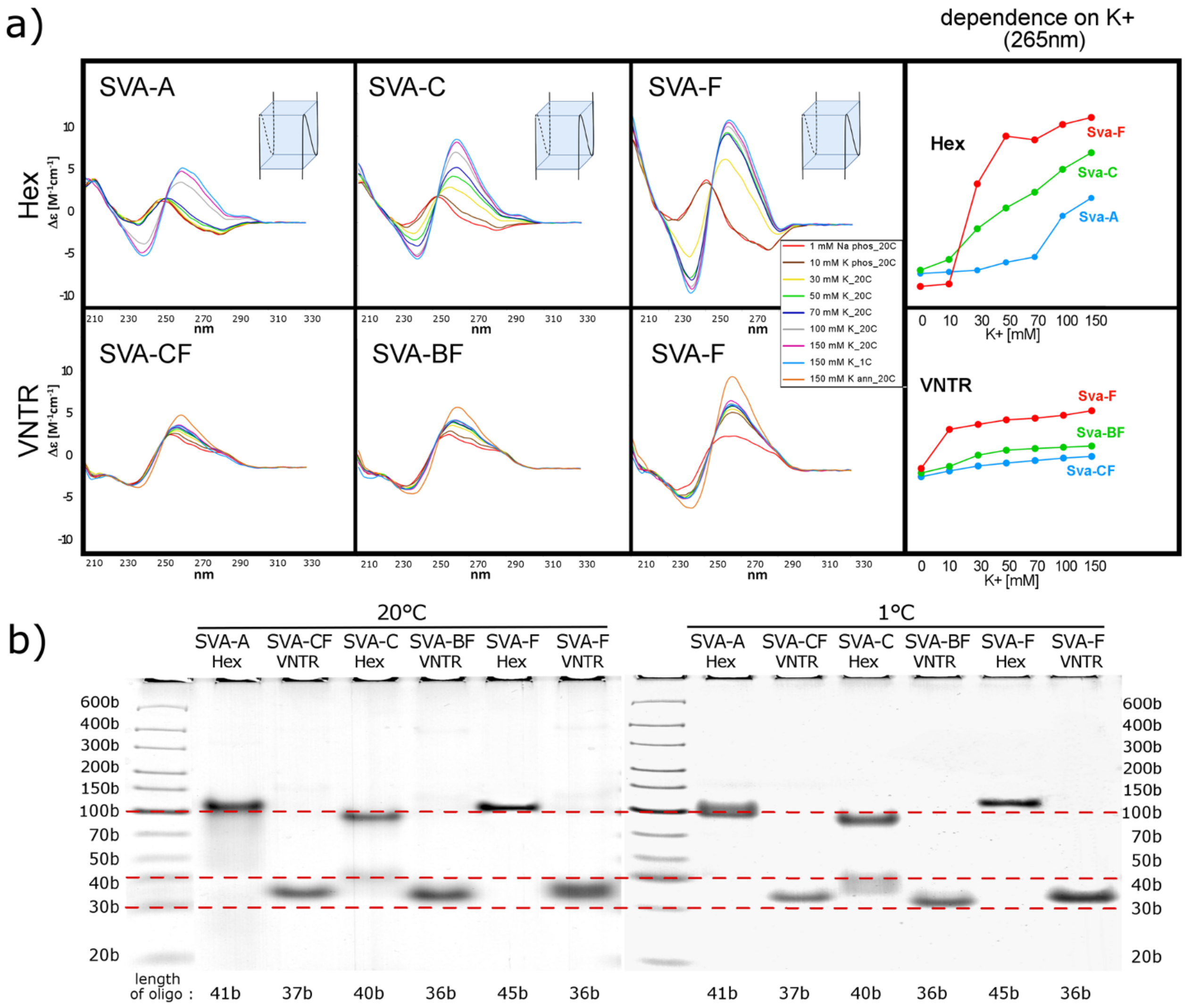


Figure 4 Circular dichroism measurements and gel electrophoreses of PQSs from SVA elements of different age. (a) CD spectra of SVA-A, SVA-C and SVA-F (shown in Table 1) obtained at various concentrations of potassium ions and at various temperatures are marked by different colors. The peak at 265 nm corresponds to a parallel-stranded quadruplex. Sketches correspond to the most probable folding of the dominating quadruplex structure according to CD and electrophoretic results. **(b)** Native gel electrophoreses of SVA PQS in the presence of 150 mM KCl at 20°C or 1°C.

minus strand (HERV_1) and the second PQS originated from the gag-pol region of the minus strand (HERV_2, Figure 1). CD measurements indicated gradual formation of parallel-stranded quadruplexes with both PQSs (Figure 5a). Native PAGE revealed that a monomolecular quadruplex structure dominated in both, HERV_1 and HERV_2 at room temperature, while a bimolecular quadruplex, in addition to two types of monomolecular ones, were formed by HERV_1 at low temperatures (Figure 5b).

In Alu elements, two PQSs (Alu-S_1 and Alu-S_2) were selected for CD measurements, both from the left part of left monomer located in plus strand (Figure 1). Both PQSs corresponded to the middle-aged Alu elements (Alu-S). Although CD spectra of both oligonucleotides indicated the formation of parallel-stranded quadruplex, the spectral changes induced by the increasing potassium concentration were gradual and limited in the case of Alu-S_1. This along with the shoulder on the long wavelength part of the positive 260nm CD band (B-DNA displays a positive maximum around 280nm) indicates that a substantial part of Alu-S_1 sequence formed a hairpin. Alu-S_2 formed the quadruplex at much lower potassium concentration (Figure 5a) and the transition was highly cooperative. The quadruplex was parallel and intramolecular in the same way as was the quadruplex of Alu-S_1 (Figure 5b). Note that the mobility of the studied quadruplexes is slower than would correspond to their length

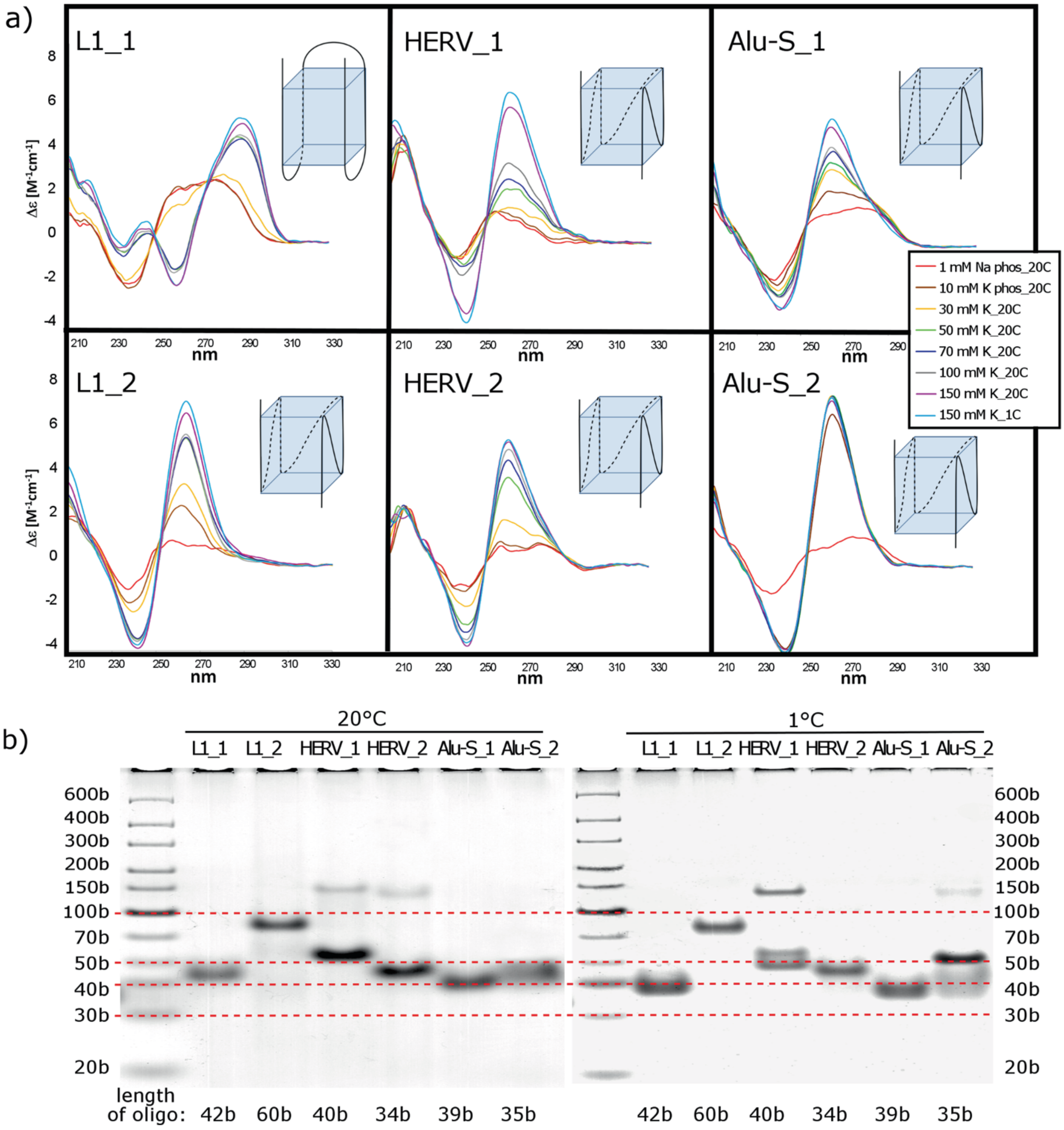


Figure 5 Circular dichroism measurements and native gel electrophoreses of PQSs from LINE-1, HERV and Alu elements. (a) CD spectra of the oligonucleotides (shown in Table S1) obtained at various concentrations of potassium and at various temperatures are marked by different colors. The peak at 265 nm indicates formation of the parallel-stranded quadruplex while maximum at 295 nm corresponds to an antiparallel-stranded quadruplex. Sketches correspond to the most probable folding of the dominating quadruplex structure according to CD and electrophoretic results. **(b)** Native gel electrophoreses of LINE-1, HERV and Alu PQS in the presence of 150 mM KCl at 20°C or 1°C.

markers. This is usually the case with the heavy G-rich strands. Moreover, the mobilities of the intramolecular qudruplexes differ (more than follows from their lengths), which may be partly a consequence of their distinct compactness, and mainly, by distinct hindering effects of the overlapping nucleotides not involved in the quadruplex structure. In addition, we measured PQSs selected from older and younger Alu families (Alu-J and Alu-Y, respectively) but we found no correlation between susceptibility to form quadruplex and the age of Alu elements.

Discussion

We found that potential quadruplex-forming sequences are located in specific regions of human transposable elements and experimentally verified the ability of such sequences to adopt quadruplex DNA conformation. Full-length and active L1 elements and younger SVA elements had a larger number of PQSs. The propensity of these sequences to form quadruplex and quadruplex stability (not shown) were higher than in older elements. Alu elements contained PQSs not inside but in their neighborhood where more PQSs were present in downstream regions of older elements.

Two available counts of G4-quadruplexes in the entire human genome found about 375,000 PQSs [24,30]. This allows us to express our numbers as proportions of mobile element PQSs to whole-genome PQS content with a value of 71%. The four main classes of elements studied here carry 49% of total predicted PQSs. These numbers reflect the current human genome sequencing and annotation status and are very likely to miss potential PQSs in centromeres, telomeres or other difficult-to-map regions of the human genome.

Our results are in agreement with Savage et al. [22] who also found that the youngest SVA (SVA-E, SVA-F) contained more quadruplexes than older elements. Such age-dependent distribution of PQSs (Figure 2) can be explained by the action of constraints leading to fixation of quadruplexes in recent and active elements while non-active older elements accumulate mutations that hinder quadruplex formation. Moreover, we found that quadruplexes are present in the central part of SVA elements in plus strand and in the left part of minus strand. If the localization of quadruplexes in plus strand has negative effect on transcription and their presence in minus strand has a positive effect [15,21], then the potential evolutionary balancing of quadruplexes abundance (an increase or a decrease) in complementary strands could regulate element activity over time.

The greater abundance of PQSs (that are GC-rich) in the neighborhood of older Alu elements is probably related to generally high GC-content of isochores containing older Alus [31]. Surprisingly, despite the age-dependent increase of GC-content of Alu neighborhood, the abundance of PQSs inside Alu elements was very low (Figure 3) and did not increase with the element age (Figure 2).

We have shown that PQSs are strongly accumulated in 3'-UTR of LINE-1 elements. Quadruplexes located in 3'-UTR can have an effect on target-primed reverse transcription (TPRT) that starts at the 3' end. Quadruplexes formed either by RNA template or by the growing first DNA strand can represent a barrier for reverse transcription. However, quadruplex DNA can regulate not only the transposable element itself but can also influence neighboring genes as was proposed recently by Kejnovsky and Lexa [21]. Because SVA elements are preferentially located inside genes or in their neighborhood [22] we suggest that recent SVA elements could spread quadruplex motifs close to genes or into genes and in this way they regulate expression of these genes. The regulatory potential of quadruplexes inside TEs decreases as the element gets older and is eroded by mutations and rearrangements. In this way, quadruplexes can enlarge the potential of transposable elements to respond to environmental challenges as was suggested by McClintock [32] long time ago.

Quadruplexes carried by TEs can also affect other cellular processes like replication or epigenetic regulation. It is remarkable that quadruplexes are located close to the LINE-1 poly(dA) tail that represents the labile region of duplex DNA. Other labile (AT-rich) regions are represented by replication origins and, surprisingly, also here quadruplexes are located [33]. Because quadruplexes also represent barriers for replication, or at least can slow it down, the spreading of PQSs by retrotransposons can also contribute to the regulation of replication speed. In addition, the quadruplexes can represent epigenetic marks in large introns that contain repetitive DNA and are also AT-rich [21,34]. Moreover, if non-B DNA conformations are nucleosome-free [35,36] and some transposable elements are preferentially inserted into naked DNA [37], then one would expect that such regions could represent sites for nested insertions, at least in some TE families.

Several proteins were shown to bind quadruplex DNA [15,38]. For example, p53 protein, that has binding sites inside human Alu and L1 elements [39,40], can strongly bind quadruplex DNA [41]. Another example is the recombination and repair protein Ku70 that was shown to bind cDNA of Ty1 yeast retrotransposons [42] and has high affinity to quadruplex DNA [43]. In this context, it is interesting that human LINEs have many Ku70/80 binding sites [44].

Taken together, the remarkable ability of some proteins to bind both TEs and quadruplex DNA underlining the relationship of these unusual DNA conformations with transposable elements as well as the higher abundance of PQSs inside younger, full-length and active elements indicates the role of quadruplexes in TE spreading. Such a role can consist in negative or positive regulation of TE activity, e.g. in response to current intracellular ionic conditions influencing the stability of quadruplexes. In the long-term perspective, quadruplexes can represent an evolutionary feedback suppressing non-controlled amplification of active elements.

Conclusions

The results suggest that activity of transposable elements, especially LINE-1 and SVA elements, contributes towards genome-wide quadruplex distribution in human. Conservation of quadruplexes at specific positions implies their function either in the life cycle of transposable elements or host genome maintenance, or both. All tested PQSs were able to form quadruplex structure in vitro, albeit with differing willingness, strand orientation and molecularity. LINE-1 and SVA families displayed an age-dependent pattern with younger elements containing a higher number of more stable quadruplexes. Further studies should be done to determine how the conserved elements are selected for during evolution.

Methods

Search for potential quadruplex-forming sequences inside transposable element

Repetitive sequences in the human genome were collected using UCSC Table Browser data [45]. The repeats from Repeat Masker track [46] (RepeatMasker, www.repeatmasker.org) from the hg38 version of the human genome were extended 200 bp in both directions and exported from Table Browser in FASTA format. The header of each sequence contained the precise position of each sequence in the hg38 assembly of the human genome, including the harboring chromosome. It also identified the class and family of element by name as returned by Repeat Masker. These identifiers were used in assigning data and results to repeats, chromosomes or to calculate whether a detected feature was inside or outside the studied repetitive region. A feature was considered to be inside only if one of its ends localized to the TE proper (not the flanking region). This dataset also includes truncated or fragmented sequences. In selected analyses, we used full-length elements, using only TEs that were longer than two thirds of a typical representative, resulting in the following thresholds [given in bp]: L1 - 4,700, Alu - 250, SVA - 1,600, HERV (ltr) - 300, HERV (internal) - 2000.

The collected sequences were scanned for the occurrence of the typical PQS3 pattern GGG-N_{1-7}-GGG-N_{1-7}-GGG-$_{1-7}$-GGG on both strands and labelled PQS3+ and PQS3-, respectively. The scan used a Perl script based on the regular expressions used in our previous study [20], recording the position and identity of each PQS3 pattern for subsequent counting and plotting. To verify that PQS frequency is not simply determined by the overall GC-content of the respective region, we calculated the expected number of PQSs in a random sequence generated by a second-order Markov model. This model was derived from the original sequence in windows of 150 bp as described previously [20,23].

CD spectroscopy and polyacrylamide gel electrophoresis

High-quality oligonucleotides (lyophilized) were purchased from Generi Biotech (Hradec Králové, Czech Republic) and dissolved in 1 mM sodium phosphate buffer with 0.3 mM EDTA (pH 7.0) to obtain final stock concentration 100 OD.ml^{-1}. Chemicals of analytical grade (Sigma-Aldrich) and deionized water (18×10^6 ohm resistance, Elga) were used for buffers. The exact oligonucleotide concentration was determined by absorbance measurements of appropriately diluted samples at 90°C in the above buffer using Unicam 5625 UV/VIS spectrophotometer and molar extinction coefficients calculated according to Gray et al. [47]. Before any measurements the DNA samples were denatured for 2 min at 90°C and slowly cooled to room temperature.

CD measurements were done using a Jasco 815 dichrograph in 1 cm Hellma cells, placed in a temperature-controlled holder. Circular dichroism was expressed as the difference in the molar absorption of the left-handed and right-handed circularly polarized light, $\Delta\epsilon$ in units of $M^{-1}cm^{-1}$. The molarities (M) were related to nucleosides. Experimental conditions were changed directly in the cells by adding concentrated solutions of potassium chloride and the final sample concentration was corrected for the volume increase. All the presented K^+ dependences were measured at 20° and 1°C.

Native polyacrylamide gel electrophoresis was performed in a temperature-controlled electrophoretic apparatus (SE-600; HoeferScientific). The gel concentration was 16% (29:1 monomer to bis ratio; Applichem). Two micrograms of oligonucleotide dissolved in 10 mM potassium phosphate and 135 mM potassium chloride were loaded into each lane. Samples were electrophoresed in 70 mM concentration of K^+ ions at 20°C for 18 h at 30V or at 1°C for 18 h at 55V. Gels were stained with Stains All (Sigma) after electrophoresis and scanned using the Personal Densitometer SI, model 375-A (Molecular Dynamics).

Additional file

Additional file 1: A detailed visualization of PQS coverage of main human transposable element families and subfamilies.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

ML, PS, BV and EK conceived the research and wrote the manuscript, PS and MV analysed quadruplexes experimentally, ML and TM carried out sequence analysis. All authors read and approved the final manuscript.

Acknowledgements

This work was supported by the Czech Science Foundation [grants P305/10/0930 to EK, P501/10/0102, P501/10/P483 to BV and P205/12/0466 to MV]; by the project "CEITEC - Central European Institute of Technology" [CZ.1.05/1.1.00/02.0068] from European Regional Development Fund; by the project OPVK [CZ.1.07/2.3.00/20.0045] and by the framework of the IT4Innovations Centre of Excellence project, [CZ.1.05/1.1.00/02.0070] supported by Operational Programme 'Research and Development for Innovations'.

Author details
[1]Faculty of Informatics, Masaryk University Brno, Botanická 68a, 60200 Brno, Czech Republic. [2]Department of Plant Developmental Genetics, Institute of Biophysics, Academy of Sciences of the Czech Republic, Královopolská 135, 61265 Brno, Czech Republic. [3]Department of Computer Systems, Faculty of Information Technology, Božetěchova 1/2, 61266 Brno, Czech Republic. [4]Department of CD Spectroscopy of Nucleic Acids, Institute of Biophysics, Academy of Sciences of the Czech Republic, Královopolská 135, 61265 Brno, Czech Republic. [5]Laboratory of CD Spectroscopy of Nucleic Acids and Proteins, CEITEC - Central European Institute of Technology, Masaryk University, Kamenice 5, 62500 Brno, Czech Republic.

Received: 18 August 2014 Accepted: 29 October 2014
Published: 27 November 2014

References

1. Gotea V, Makalowski W: **Do transposable elements really contribute to proteomes?.** *Trends Genet* 2006, **22:**260–261.
2. Britten R: **Transposable elements have contributed to thousands of human proteins.** *Proc Natl Acad Sci USA* 2006, **103:**1798–1803.
3. Kazazian JHH, Wong C, Youssoufian H, Scott AF, Phillips DG, Antonarakis SE: **Haemophilia a resulting from de novo insertion of l1 sequences represents a novel mechanism for mutation in man.** *Nature* 1988, **332:**164–166.
4. Miki Y, Nishisho I, Horii A, Miyoshi Y, Utsunomiya J, Kinzler KW, Vogelstein B, Nakamura Y: **Disruption of the apc gene by a retrotransposal insertion of l1 sequence in colon cancer.** *Cancer Res* 1992, **52:**643–645.
5. Bailie JK, Barnett MW, Upton KR, Gerhardt DJ, Richmond TA, De Sapio F, Brennan PM, Rizzu P, Smith S, Fell M, Talbot RT, Gustincich S, Freeman TC, Mattick JS, Hume DA, Heutink P, Carninci P, Jeddeloh JA, Faulkner GJ: **Somatic retrotransposition alters the genetic landscape of the human brain.** *Nature* 2011, **479:**534–537.
6. Evrony GD, Cai X, Lee E, Hills LB, Elhosary PC, Lehmann HS, Parker JJ, Atabay KD, Gilmore EC, Poduri A, Park PJ, Walsh CA: **Single-neuron sequencing analysis of l1 retrotransposition and somatic mutation in the human brain.** *Cell* 2012, **151:**483–496.
7. Lee E, Iskow R, Yang L, Gokcumen O, Gokcumen O, Haseley P, Luquette LJr, Lohr JG, Harris CC, Ding L, Wilson RK, Wheeler DA, Gibbs RA, Kucherlapati R, Lee C, Kharchenko PV, Park PJ: **Landscape of somatic retrotransposition in human cancers.** *Science* 2012, **337:**967–971.
8. Babatz TD, Burns KH: **Functional impact of the human mobilome.** *Curr Opin Genet Dev* 2013, **23:**264–270.
9. Cordaux R, Batzer MA: **The impact of retrotransposons on human genome evolution.** *Nat Rev Genet* 2009, **10:**691–703.
10. Biemont C, Vieira C: **Junk dna as an evolutionary force.** *Nature* 2006, **443:**521–524.
11. Mayer J, Meese E: **Human endogenous retroviruses in the primate lineage and their influence on host genomes.** *Cytogenet Genome Res* 2005, **110:**448–456.
12. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, Funke R, Gage D, Harris K, Heaford A, Howland J, Kann L, Lehoczky J, LeVine R, McEwan P, McKernan K, Meldrim J, Mesirov JP, Miranda C, Morris W, Naylor J, Raymond C, Rosetti M, Santos R, Sheridan A, Sougnez C, *et al.*: **Initial sequencing and analysis of the human genome.** *Nature* 2001, **409:**860–921.
13. Wang H, Xing J, Grover D, Hedges DJ, Han K, Walker JA, Batzer MA: **Sva elements: a hominid-specific retroposon family.** *J Mol Biol* 2005, **354:**994–1007.
14. Mills RE, Bennett EA, Iskow RC, Devine SE: **Which transposable elements are active in the human genome?** *Trends Genet* 2007, **23:**183–191.
15. Bochman M. J, Paeschke K, Zakian VA: **Dna secondary structures: stability and function of g-quadruplex structures.** *Nat Rev Genet* 2012, **13:**770–780.
16. Sundquist WI, Heaphy S: **Evidence for intrastrand quadruplex formation in the dimerization of human immunodeficiency virus 1 genomic rna.** *Proc Natl Scad Sci USA* 1993, **90:**3393–3397.
17. Howell R, Usdin K: **The ability to form intrastrand tetraplexes is an evolutionary conserved feature of the 3′ end of l1 retrotransposons.** *Mol Biol Evol* 1997, **14:**144–155.
18. Nambiar M, Goldsmith G, Moorthy BT, Lieber MR, Joshi MV, Choudhary B, Hosur RV, Raghavan SC: **Formation of a q-quadruplex at the bcl2 major breakpoint region of the t(14;18) translocation in follicular lymphoma.** *Nucleic Acids Res* 2011, **39:**936–948.
19. Kapitonov VV, Jurka J: **Rag1 core and v(d)j recombination signal sequences were derived from transib transposons.** *PLoS Biol* 2005, **3:**181.
20. Lexa M, Kejnovsky E, Steflova P, Konvalinova H, Vorlickova M, Vyskot B: **Quadruplex-forming sequences occupy discrete regions inside plant ltr retrotransposons.** *Nucleic Acids Res* 2014, **42:**968–978.
21. Kejnovsky E, Lexa M: **Quadruplex-forming dna sequences spread by retrotransposons may serve as genome regulators.** *Mobile Genet Elements* 2014, **4:**101.
22. Savage AL, Bubb VJ, Breen G, Quinn JP: **Characterization of the potential function of sva retrotransposons to modulate gene expression patterns.** *BMC Evol Biol* 2013, **13:**101.
23. Huppert JL, Balasubramanian S: **Q-quadruplexes in promoters throughout the human genome.** *Nucl Acids Res* 2005, **35:**406–413.
24. Huppert JL, Balasubramanian S: **Prevalence of quadruplexes in the human genome.** *Nucl Acids Res* 2007, **33:**2908–2916.
25. Lam EYN, Beraldi D, Tannahill D, Balasubramanian S: **G-quadruplex structures are stable and detectable in human genomic dna.** *Nat Commun* 2013, **4:**1796.
26. Mills RE, Bennett EA, Iskow RC, Luttig CT, Tsui C, Pittard WS, Devine SE: **Recently mobilised transposons in the human and chimpanzee genomes.** *Am J Hum Genet* 2006, **78:**671–679.
27. Bennett EA, Keller H, Mills RE, Schmidt S, Moran JV, Weichenrieder O, Devine SE: **Active alu retrotransposons in the human genome.** *Genome Res* 2008, **18:**1875–1883.
28. Vorlickova M, Kejnovska I, Sagi J, Renciuk D, Bednarova K, Motlova J, Kypr J: **Circular dichroism and guanine quadruplexes.** *Methods* 2012, **57:**64–75.
29. Kypr J, Kejnovska I, Renciuk D, Vorlickova M: **Circular dichroism and conformational polymorphism of dna.** *Nucl Acids Res* 2009, **37:**1713–1725.
30. Todd AK, Johnstone M, Neidle S: **Highly prevalent putative quadruplex sequence motifs in human dna.** *Nucl Acids Res* 2005, **33:**2901–2907.
31. Eyre-Walker A, Hurst LD: **The evolution of isochores.** *Nat Rev Genetics* 2001, **2:**549–555.
32. McClintock B: **The significance of response of the genome to challenge.** *Science* 1983, **226:**792–801.
33. Cayrou C, Coulombe P, Puy A, Rialle S, Kaplan N, Segal E, Mechali M: **New insights into replication origin characteristics in metazoans.** *Cell Cycle* 2012, **11:**658–667.
34. Gelfman S, Cohen N, Yearim A, Ast G: **Dna-methylation effect on cotranscriptional splicing is dependent on gc architecture of the exon-intron structure.** *Genome Res* 2013, **23:**789–799.
35. Wong HM, Huppert JL: **Stable g-quadruplexes are found outside nucleosome-bound regions.** *Mol Biosyst* 2009, **5:**1713–1719.
36. De S, Michor F: **Dna secondary structures and epigenetic determinants of cancer genome evolution.** *Nat Struct Mol Biol* 2011, **18:**950–956.
37. Gangadharan S, Mularoni L, Fain-Thornton J, Wheelan SJ, Craig NL: **Dna transposon hermes inserts into dna in nucleosome-free regions in vivo.** *Proc Natl Acad Sci USA* 2010, **107:**21966–21972.
38. Whitehouse I, Owen-Hughes T: **Atrx: put me on repeat.** *Cell* 2010, **143:**335–336.
39. Cui F, Sirotkin MV, Zhurkin VB: **Impact of alu repeats on the evolution of human p53 binding sites.** *Biol Direct* 2011, **6:**2.
40. Harris CR, DeWang A, Zupnick A, Normart R, Gabriel A, Prives C, Levine AJ, Hoh J: **p53 responsive elements in human retrotransposons.** *Oncogene* 2009, **28:**3857–3865.
41. Quante T, Otto B, Brazdova M, Kejnovska I, Deppert W, Tolstonog GV: **Mutant p53 is a transcriptional co-factor that binds to g-rich regulatory regions of active genes and generates transcriptional plasticity.** *Cell Cycle* 2012, **11:**3290–3303.
42. Dawns JA, Jackson SP: **Involvement of dna end-binding protein ku in ty element retrotransposition.** *Mol Cell Biol* 1999, **19:**6260–6268.
43. Paramasivan M, Membrino A, Cogoi S, Fukuda H, Nakagama H, Xodo LE: **Protein hnrnp a1 and its derivate up1 unfold quadruplex dna in the human kras promoter: implications for transcription.** *Nucl Acids Res* 2009, **37:**2841–2853.

44. Katz DJ, Beer MA, Levorse JM, Tilghman SM: **Functional characterization of a novel ku70/80 pause site at the h19/igf2 imprinting control region.** *Mol Cell Biol* 2005, **25:**3855–3863.
45. Karolchik D, Hinrichs AS, Furey TS, Roskin KM, Sugnet CW, Haussler D, Kent WJ: **The ucsc table browser data retrieval tool.** *Nucleic Acids Res* 2004, **32**(Database issue):493–496.
46. Karolchik D, Barber GP, Casper J, Clawson H, Cline MS, Diekhans M, Dreszer TR, Fujita PA, Guruvadoo L, Haeussler M, Harte RA, Heitner S, Hinrichs AS, Learned K, Lee BT, Li CH, Raney BJ, Rhead B, Rosenbloom KR, Sloan CA, Speir ML, Zweig AS, Haussler D, Kuhn RM, Kent WJ: **The ucsc genome browser database: 2014 update.** *Nucleic Acids Res* 2014, **42:**764–770.
47. Gray DM, Hung SH, Johnson KH: **Absorption and circular dichroism spectroscopy of nucleic acid duplexes and triplexes.** *Methods Enzymol* 1995, **246:**19–34.

doi:10.1186/1471-2164-15-1032
Cite this article as: Lexa *et al.*: Guanine quadruplexes are formed by specific regions of human transposable elements. *BMC Genomics* 2014 **15**:1032.