Workshop on Bioinformatics and Computational Biology WBCB 2023

Conference Abstracts

23rd Conference ITAT
(Information Technologies – Applications and Theory)
Hotel Hutník I, Tatranské Matliare, Slovakia
September 22-26, 2023

Workshop on Bioinformatics and Computational Biology WBCB 2023
Conference Abstracts
Editors: Bronislava Brejová, Matej Lexa
Publisher: Knižničné a edičné centrum, Fakulta matematiky, fyziky a informatiky,
Univerzita Komenského, Mlynská dolina, 842 48 Bratislava
First edition, 2023

ISBN 978-80-8147-132-2

Preface

The Workshop on Bioinformatics and Computational Biology was held at Hotel Hutnik in Tatranské Matliare, Vysoké Tatry, Slovakia on September 22 – 23, 2023 as part of the ITAT conference. While the conference is a traditional forum for Slovak and Czech computer scientists, the workshop follows a more recent tradition of a relatively irregular extension of topics into the realm of bioinformatics and computational biology. Similarly to the main conference (ITAT), it serves as a platform for the exchange of ideas in a friendly but stimulating environment. The workshop allows younger researchers to meet their more seasoned colleagues in an informal setting, sparking discussions that would be hard to initiate elsewhere. For example, the settings often allow for a short hiking trip in the Slovak mountains. This year the organizers secured a keynote speaker from North America and also reached out with invitations to colleagues from neighboring countries, giving the workshop a more international tone.

The topics of the workshop included algorithmic approaches and data structures for sequencing data analysis and pangenome representation, molecular modeling, bioinformatics of repetitive sequences, and medical bioinformatics.

The workshop program consisted of an invited talk by Travis Gagie (Dalhousie University, Canada), entitled "Flexible grammar-based indexes", and 14 presentations based on 5 proceedings articles and 9 short abstracts. All proceedings papers were reviewed by at least three anonymous reviewers and their full versions are published in the proceedings of ITAT 2023 published by CEUR-WS. This volume contains abstracts of all presentations from the workshop.

September 2023

Bronislava Brejová, Matej Lexa
Program committee chairs

The WBCB 2023 Program Committee

Broňa Brejová, Comenius University in Bratislava, co-chair
Matej Lexa, Masaryk University, Brno, co-chair
Vladimir Boža, Comenius University in Bratislava
Jaroslav Budiš, Geneton, Bratislava
Krisztian Buza, Eötvös Loránd University
Vojtěch Bystrý, Central European Institute of Technology, Brno
Luca Denti, University of Milano-Bicocca
Norbert Dojer, University of Warsaw
Askar Gafurov, Comenius University in Bratislava
Mária Lucká, Comenius University in Bratislava
Szabolcs Makai, Semmelweis University
Jan Oppelt, University of Pennsylvania
Karol Pál, Pennsylvania State University
Agnieszka Rybarczyk, Poznan University of Technology
Tomáš Vinař, Comenius University in Bratislava
Martina Višňovská, Oslo University Hospital
Filip Železný, Czech Technical University in Prague

Contents

Keynote abstract

Abstracts of Proceedings Articles

Contributed Abstracts

WBCB 2023, p. 1, ISBN 978-80-8147-132-2

Flexible grammar-based indexes

Travis Gagie

Dalhousie University, Halifax, Nova Scotia, Canada

For highly repetitive texts such as pangenomic databases, indexes based on grammars (or Lempel-Ziv parses, string attractors, etc.) are usually significantly smaller than indexes based on the Burrows-Wheeler Transform. Nevertheless, they are not widely used in practice, probably because fast implementations are complicated and support only exact pattern matching. In this talk we will review theoretical and practical advances in the past ten years towards computing matching statistics, MEMs, k-MEMs and MUMs, for example, quickly with grammar-based indexes. We will close with a survey of some future advantages of challenges of using grammar-based indexes for pangenomics.

WBCB 2023, p. 2, ISBN 978-80-8147-132-2

Prefix-free graphs and suffix array construction in sublinear space

Andrej Baláž* Alessia Petescia

Department of Applied Informatics, Faculty of Mathematics, Physics and Informatics, Comenius University, Bratislava, Slovakia

A recent paradigm shift in bioinformatics from a single reference genome to a pangenome brought with it several graph structures. These graph structures must implement operations, such as efficient construction from multiple genomes and read mapping. Read mapping is a well-studied problem in sequential data, and, together with data structures such as suffix array and Burrows-Wheeler transform, allows for efficient computation. Attempts to achieve comparatively high performance on graphs bring many complications since the common data structures on strings are not easily obtainable for graphs. In this work, we introduce prefix-free graphs, a novel pangenomic data structure; we show how to construct them and how to use them to obtain well-known data structures from stringology in sublinear space, allowing for many efficient operations on pangenomes.

*andrejbalaz001@gmail.com

WBCB 2023, p. 3, ISBN 978-80-8147-132-2

Precise Nanopore Signal Modeling Improves Unsupervised Single-Molecule Methylation Detection

Vladimír Boža[1] Eduard Batmendijn Peter Perešíni[1] Viktória Hodorová[2]
Hana Lichancová[2] Rastislav Rabatin Broňa Brejová[1] Jozef Nosek[2] Tomáš Vinař[1]

[1] Faculty of Mathematics, Physics and Informatics, Comenius University in Bratislava, Mlynská dolina, 842 48 Bratislava, Slovakia
[2] Faculty of Natural Sciences, Comenius University in Bratislava, Ilkovičova 6, 841 05 Bratislava, Slovakia

Base calling in nanopore sequencing is a difficult and computationally intensive problem, typically resulting in high error rates. In many applications of nanopore sequencing, direct analysis of raw signal is a viable alternative. Dynamic time warping (DTW) is an important building block for raw signal analysis. In this paper, we propose several improvements to DTW class of algorithms to better account for specifics of nanopore signal modeling. We have implemented these improvements in a new signal-to-reference alignment tool Nadavca. We demonstrate that Nadavca alignments improve unsupervised methylation detection over Tombo. We also demonstrate that by providing additional information about the discriminative power of positions in the signal, an otherwise unsupervised method can approach the accuracy of supervised models.

Availability and implementation: Nadavca is available under MIT license at `https://github.com/fmfi-compbio/nadavca`. Nanopore sequencing data sets are available from ENA bioproject PRJEB64246. *Jaminaea angkorensis* reference genome assembly is available from Zenodo `https://doi.org/10.5281/zenodo.8145315`.

WBCB 2023, p. 4, ISBN 978-80-8147-132-2

A Simple and Effective Classifier for the Detection of Psychotic Disorders based on Heart Rate Variability Time Series

Krisztian Buza[1,2]* Kamil Książek[3] Wilhelm Masarczyk[4] Przemysław Głomb[3]
Piotr Gorczyca[4] Magdalena Piegza[4]

[1] Artificial Intelligence Laboratory, Institute Jozef Stefan, Jamova cesta 39, 1000 Ljubljana, Slovenia
[2] BioIntelligence Group, Department of Mathematics-Informatics, Sapientia Hungarian University of Transylvania, Targu Mures, Romania
[3] Institute of Theoretical and Applied Informatics, Polish Academy of Sciences, Bałtycka 5, 44-100 Gliwice, Poland
[4] Department of Psychiatry, Faculty of Medical Sciences in Zabrze, Medical University of Silesia, Pyskowicka 49, 42-612 Tarnowskie Góry, Poland

In this paper, we focus on automated detection of schizophrenia and bipolar disorder. For this task, we describe a simple and effective classifier, i.e. convolutional nearest neighbor. It provides a data-driven and objective approach for the detection of schizophrenia and bipolar disorder based on heart rate variability time series. According to our results, our approach is able to distinguish whether the selected person belongs to the patient group with an accuracy of 85% and area under receiver-operator characteristic curve of 0.92.

*buza@biointelligence.hu

WBCB 2023, p. 5, ISBN 978-80-8147-132-2

ESGq: Alternative Splicing events quantification across conditions based on Event Splicing Graphs

Davide Cozzi* Paola Bonizzoni Luca Denti

Department of Informatics, Systems and Communication, University of Milano-Bicocca, Milan, Italy

Alternative Splicing (AS) is a regulation mechanism that contributes to protein diversity and is also associated to many diseases and tumors. Alternative splicing events quantification from RNA-Seq reads is a crucial step in understanding this complex biological mechanism. However, tools for AS events detection and quantification show inconsistent results. This reduces their reliability in fully capturing and explaining alternative splicing. We introduce `ESGq`, a novel approach for the quantification of AS events across conditions based on read alignment against Event Splicing Graphs. By comparing `ESGq` to two state-of-the-art tools on real RNA-Seq data, we validate its performance and evaluate the statistical correlation of the results. `ESGq` is freely available at https://github.com/AlgoLab/ESGq.

*d.cozzi@campus.unimib.it

WBCB 2023, p. 6, ISBN 978-80-8147-132-2

Identifying Clusters in Graph Representations of Genomes

Eva Herencsárová Broňa Brejová

Department of Computer Science, Faculty of Mathematics, Physics and Informatics, Comenius University, Bratislava, Slovakia

In many bioinformatics applications the task is to identify biologically significant locations in an individual genome. In our work, we are interested in finding high-density clusters of such biologically meaningful locations in a graph representation of a pangenome, which is a collection of related genomes. Different formulations of finding such clusters were previously studied for sequences. In this work, we study an extension of this problem for graphs, which we formalize as finding a set of vertex-disjoint paths with a maximum score in a weighted directed graph. We provide a linear-time algorithm for a special class of graphs corresponding to elastic-degenerate strings, one of pangenome representations. We also provide a fixed-parameter tractable algorithm for directed acyclic graphs with a special path decomposition of a limited width.

WBCB 2023, p. 7, ISBN 978-80-8147-132-2

Study of conformational changes of Tau(210-240) upon multiple phosphorylations using molecular dynamics simulations

Krishnendu Bera[123*] Thomas Fellmeth[12†] Alessia Lasorsa[45‡] Isabelle Landrieu[45§]
Jozef Hritz[13¶]

[1] CEITEC MU, Masaryk University, Kamenice 753/5, 625 00 Brno, Czech Republic
[2] NCBR, Faculty of Science, Masaryk University, Kamenice 5, 625 00, Brno, Czech Republic
[3] Deptartment of Chemistry, Faculty of Science, Masaryk University, Kamenice 5, 625 00 Brno, Czech Republic
[4] CNRS EMR9002 Integrative Structural Biology, F-59000, Lille, France
[5] Univ. Lille, Inserm, CHU Lille, Institut Pasteur de Lille, U1167 - RID-AGE - Risk Factors and Molecular Determinants of Aging-Related Diseases, F-59000, Lille, France

It is challenging to elucidate the conformational dynamics of intrinsically disordered proteins (IDPs) regulated by post-translational modifications (PTMs) such as phosphorylation. Tau is a well-known IDP, found hyperphosphorylated in Alzheimer's disease (AD) in humans [3]. The proline-rich domain of tau directly interacts with its partner proteins such as BIN1, 14-3-3 etc. All atoms molecular dynamic (MD) simulation studies have been performed in microsecond time scale for wildtype and four phosphorylated (pT212, pT217, pT231, pS235) tau(210-240) peptide using three different temperatures (278K, 298K and 310K) and two different force field parameters (AMBER99SB-ILDN and CHARMM36m) with TIP4PD water model as combination of these parameters worked the better for IDPs found from our group previous studies [4, 2]. These four-phosphorylations cause increase in compactness of the peptide resulting bent conformation. From the experimental studies we found the binding affinity reduced by 12-folds between SH3 domain of BIN1 protein and tau(210-240). The binding of associated proteins like BIN1 with tau may alter by the strong salt bridges, forming nearby lysine and arginine due to the phosphorylation [1]. Phosphorylation induces a strong structural transition, with tau(210-240) favouring a bent conformation. Currently we are testing the coarse grain force fields. The MD simulation results were verified using NMR experimental parameters like chemical shift and ^{3}J-coupling [1].

Acknowledgements. Computational resources were supplied by the metacentrum and IT4 Innovations National Supercomputing Center (OPEN-17-7), project (e-INFRA CZ 90140 and LM2018140) provided within the program Projects of Large Research, Development and Innovations Infrastructures. KB and JH were funded by the Ministry of Education, Youth, and Sport of the Czech Republic (MEYS CR), grant number LTAUSA18168 (Inter-Excellence Inter-Action), Czech Science Foundation [GF20-05789L] and the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement no. 873127 – InterTAU. KB is also supported by Brno Ph.D. Talent Scholarship – funded by the Brno City Municipality, Brno, Czech Republic.

References

[1] A. Lasorsa, K. Bera, Idir Malki, Elian Dupré, F.-X. Cantrelle, H. Merzougui, D. Sinnaeve, X. Hanoulle, J. Hritz, and I. Landrieu. Conformation and affinity modulations by multiple phosphorylation occurring in the bin1 sh3 domain binding site of the tau protein proline-rich region. *Biochemistry*, 62(11):1631–1642, 2023.

[2] E. Lucendo, M. Sancho, F. Lolicato, M. Javanainen, W. Kulig, D. Leiva, G. Duarte, V. Andreu-Fernández, I. Mingarroe, and M. Orzáez. Mcl-1 and bok transmembrane domains: Unexpected players in the modulation of apoptosis. *PNAS*, 117(1):27980–27988, 2020.

[3] T. Shimada, A. E. Fournier, and K. Yamagata. Neuroprotective function of 14-3-3 proteins in neurodegeneration. *BioMed research international*, 2013.

[4] V. Zapletal, A. Mládek, K. Melková, P. Louša, E. Nomilner, Z. Jaseňáková, V. Kubáň, M. Makovická, Laníková, L. Žídek L, and J. Hritz. hoice of force field for proteins containing structured and intrinsically disordered regions. *Biophys J*, 118(7):1621–1633, 2020.

[*] Krishnendu.Bera@ceitec.muni.cz
[†] Thomas.Fellmeth@ceitec.muni.cz
[‡] isabelle.landrieu@univ-lille.fr
[§] alessia.lasorsa@gmail.com
[¶] Jozef.Hritz@ceitec.muni.cz

WBCB 2023, p. 8, ISBN 978-80-8147-132-2

de Bruijn graph and variation graph - common axiomatization of pangenome models

Adam Cicherski Norbert Dojer

University of Warsaw, Warsaw, Poland

Pangenomes serve as a fundamental framework for collectively studying the genomes of closely related organisms. Several pangenome models have been introduced, each offering distinct functionalities and applications through available tool. Notably, among the graph-based models, variation graphs and de Bruijn graphs are widely embraced and utilized. In our recently written paper [1] we investigated the relationship between these two models.

De Bruijn graphs (dBGs) consist of nodes uniquely labeled with k-mers, while edges represent overlaps of length $k-1$ between these $k-$mers. The construction of a dBG for a specific set of genomes is a straightforward process and can be accomplished in linear time due to the strict determination by the parameter k.

Variation graphs (VG) have nodes labeled with DNA sequences of arbitrary length. Genomic sequences are represented in the graph by paths, for which the concatenation of labels form the respective sequences. Such structure allows to avoid the redundancy of the dBG representation however The construction of VG models is more computationally resource-intensive and there are many possible variation graphs that represent a particular collection of genomes. Two completely uninformative extremes are: an empty graph with each node labeled with one of the genome sequences, and a graph with 4 nodes labeled with single letters.

In our work [1], we proposed an axiomatization of the desirable properties inherent to a graph representation of a set of strings.

We introduced two crucial attributes for this representation: *k-completeness* and *k-faithfulness*. Let G be a VG and π a set of paths representing set of input strings S. In simple terms, the representation is deemed k-complete if every k-mer in S is depicted by the same path in the graph, and is k-faithful if all multiple occurrences of the same vertex in paths representing genome are essential to satisfy k-completeness. To be more precise, the concept of k-faithfulness is further elaborated through the following relationships:

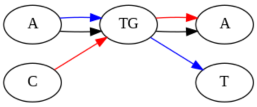

Figure 1: Example of 3-complete and 3−faithful representation of set of strings ATGA, CTGA ATGT. Each 3-mer is represented by unique path. Occurrences of vertex TG in blue and black path are directly 3-extendable to the path labeled by ATG and occurences on red and black path are directly 3−extendable to the path labeled by TGA.

- The pair of occurrences of a vertex v is *directly k-extendable* if both of these occurrences are encompassed within a shared subpath labeled with a string of length $\geq k$.
- The pair of occurrences of a vertex v is *k-extendable* if there is a sequence of occurrences of v that includes both occurrences from this pair and each two consecutive elements in that sequence are directly k-extendable.

Consequently, we defined the representation as *k-faithful* if every pair of occurrences of a vertex satisfies the condition of being k-extendable.

We proved the VG satisfying both these criteria exist for every input set of strings. We showed the uniqueness of such graph up to splitting of non branching paths. Moreover, we described the algorithm for building of such VG from dBG. Additionally, it facilitates the seamless transfer of annotations between both models and enables comparative analyses of results derived from each respective model.

References

[1] Adam Cicherski and Norbert Dojer. From de bruijn graphs to variation graphs – relationships between pangenome models. June 2023. SPIRE2023: 30th International Symposium on String Processing and Information Retrieval.

WBCB 2023, p. 9, ISBN 978-80-8147-132-2

Wheeler maps

Adrián Goga[1]

joint work with

Andrej Baláž[2], Travis Gagie[3], Simon Heumos[4],
Gonzalo Navarro[5], Alessia Petescia[2] and Jouni Sirén[6]

[1] Department of Computer Science, Comenius University in Bratislava, Bratislava, Slovakia
[2] Department of Applied Informatics, Comenius University in Bratislava, Bratislava, Slovakia
[3] Faculty of Computer Science, Dalhousie University, Halifax, NS, Canada
[4] Quantitative Biology Center (QBiC), University of Tübingen, Tübingen, Germany
[5] Department of Computer Science, University of Chile, Santiago de Chile, Chile
[6] Genomics Institute, University of California, Santa Cruz, CA, USA

Motivated by the challenges in pangenomic read alignment, we propose a generalization of Wheeler graphs [2] that we call Wheeler maps. A Wheeler map stores a text $T[1..n]$ and an assignment of tags to the characters of T such that we can preprocess a pattern $P[1..m]$ and then, given i and j, quickly return all the distinct tags labelling the first characters of the occurrences of $P[i..j]$ in T.

For the applications that most interest us, characters with long common contexts are likely to have the same tag, so we consider the number t of runs in the list of tags sorted by their characters' positions in the Burrows-Wheeler Transform (BWT) of T. We show how, given a straight-line program with g rules for T, we can build an $O(g+r+t)$-space Wheeler map, where r is the number of runs in the BWT of T, with which we can preprocess a pattern $P[1..m]$ in $O(m \log n)$ time and then return the k distinct tags for $P[i..j]$ in the optimal $O(k)$ time for any given i and j. To this end, we combine the r-index machinery [4] for compressed text indexing with the document listing data structure of Muthukrishnan [3].

Furthermore, for a parameter f fixed at construction time, we show how we can efficiently report all the distinct tags that each label at least f occurrences of $P[i..j]$ in T. In addition, we also provide techniques to efficiently count the number of distinct tags of $P[i..j]$ and to list top-k most frequent tags that label the occurrences of $P[i..j]$.

Our work provides a convenient middle ground between the problem of pattern matching on labeled graphs, for which Equi et al. [1] showed it cannot run in sub-quadratic time, and compressed text indexing, which is regarded as one of the success stories in the field. In particular, our work allows efficient locating of maximal exact matches (MEMs) – a popular choice for seeds in read alignment – to work on pangenome graphs. Apart from most of the previous work in the field, we do not place any restrictions on the graph topology. The main virtue of our method is that we do not discard any genomic variations present in the data, nor do we introduce any chimeric variations that are absent in the data.

Acknowledgements. This work has been supported by the Comenius University grant for doctoral students No. 406.

References

[1] Massimo Equi, Veli Mäkinen, Alexandru I Tomescu, and Roberto Grossi. On the complexity of string matching for graphs. *ACM Transactions on Algorithms*, 19(3):1–25, 2023.

[2] Travis Gagie, Giovanni Manzini, and Jouni Sirén. Wheeler graphs: A framework for BWT-based data structures. *Theoretical computer science*, 698:67–78, 2017.

[3] Shanmugavelayutham Muthukrishnan. Efficient algorithms for document retrieval problems. In *Proc. SODA*, pages 657–666, 2002.

[4] Massimiliano Rossi, Marco Oliva, Ben Langmead, Travis Gagie, and Christina Boucher. MONI: a pangenomic index for finding maximal exact matches. *Journal of Computational Biology*, 29(2):169–187, 2022.

WBCB 2023, p. 10, ISBN 978-80-8147-132-2

Analysis and classification of long terminal repeat sequences from plant LTR-retrotransposons - Abstract

Jakub Horváth[1]* Matej Lexa[1]†

Faculty of Informatics, Masaryk University, Botanická 68A,
602 00 Brno-Královo Pole, Czech Republic

Long Terminal Repeats (LTRs) are repetitive DNA sequences widely distributed throughout eukaryotic genomes, found in particular abundance in retrotransposons. A deeper understanding of the structural and functional characteristics of LTRs is crucial for deciphering their role in genome evolution and their involvement in disease mechanisms.

This research presents a comprehensive analysis of frequently co-occurring motifs within LTRs and employs increasingly complex classification methods for identifying LTR sequences. Specifically, this study explores the use of a gradient boosting classifier, a neural network model that combines convolutional neural networks (CNNs) and Long Short-Term Memory (LSTM) networks, and an innovative approach that uses a fine-tuned version of the DNABERT[2] transformer architecture model called LTRBERT[1].

In order to gain a deeper insight into the structure of LTRs and avoid treating the created models as black boxes, we employed several model interpretability techniques, with the aim of uncovering significant regions and features within these sequences.

Acknowledgements. This work has been supported by the Grant Agency of the Czech Republic grant number GA21-00580S.

References

[1] Jakub Horváth. Analysis and classification of long terminal repeat (ltr) sequences using machine learning approaches. Master's thesis, 2023. URL https://is.muni.cz/th/m8eg3/.

[2] Yanrong Ji, Zhihan Zhou, Han Liu, and Ramana V Davuluri. DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome. *Bioinformatics*, 37(15): 2112–2120, 02 2021. ISSN 1367-4803. doi: 10.1093/bioinformatics/btab083. URL https://doi.org/10.1093/bioinformatics/btab083.

*jakubhorvath119@gmail.com

†lexa@fi.muni.cz

WBCB 2023, p. 11, ISBN 978-80-8147-132-2

In silico designing of Insecticides Against *Bemisia tabaci* targeting ecdysone receptor

Indu[45*] Jozef Hritz[12†] Václav Brázda[47‡] Rajesh Kumar[6§] Krishnendu Bera[123¶]

[1] Central European Institute of Technology, Masaryk University, Brno, Czech Republic
[2] Department of Chemistry, Faculty of Science, Masaryk University, Kamenice 5, 625 00 Brno, 17 Czech Republic
[3] National Centre for Biomolecular Research, Faculty of Science, Masaryk University, Brno, Czech Republic
[4] Institute of Biophysics of the Czech Academy of Sciences, Brno, Czech Republic
[5] Department of Experimental Biology, Faculty of Science, Masaryk University, Brno, Czech Republic
[6] Plant Biotechnology Laboratory, Department of Genetics and Plant Breeding, Rajiv Gandhi South Campus (RGSC), Banaras Hindu University, Mirzapur 231001, India
[7] Brno University of Technology, Faculty of Chemistry, Brno, Czech Republic

The Bemisia tabaci commonly known as white fly, is one of the major destructive pest which destroys more than 600 crop species worldwide. It carries more than 100 viruses in plants which interferes with plant growth. This project aims to find novel lead molecules using in silico approaches. The protein ecdysone receptor (EcR) of B. tabaci is involved in reproduction processes, metamorphosis and cell differentiation. There is no homologous protein is found in mammals which makes it an ideal insecticide target. There is no full-length experimental structure in PDB. We modelled the full-length protein using Alphafold 2.2.0 [3]. The intrinsic disordered regions were predicted using IDP predictor software, i.e. DEPICTER [1]. Further, we retrieved 32,552 bacterial and fungal secondary metabolites from the npatlas 2.0 database [5] and then we docked each metabolite with the MD simulated obtained last conformation of EcR protein using idock 2.2.3 software [4]. We have set a cut-off -10 kcal/mol binding energy and found 14 metabolites. We have redock these 14 metabolites again with Autodock vina 1.1.2 [2] to validate idock 2.2.3 results and found an almost similar result with minor deviations. These dockings were compared with 20E, a natural hormone binding with EcR protein. Lastly, one compound K6323 with the most suitable scoring function were selected for 30 ns MD simulations of the protein complex with E20 and K6323 and compared with apo form of the protein. Further, QMMM/GBSA-based binding energy was calculated from 100 snapshots from MD simulation. The binding energy of K6323 was found to be better than the natural inhibitor 20E. These computational predictions can be analyzed further experimentally.

Acknowledgements. Computational resources were supplied by the metacentrum and IT4 Innovations National Supercomputing Center (OPEN-17-7), project (e-INFRA CZ 90140 and LM2018140) provided within the program Projects of Large Research, Development and Innovations Infrastructures. KB and JH were funded by the Ministry of Education, Youth, and Sport of the Czech Republic (MEYS CR), grant number LTAUSA18168 (Inter-Excellence Inter-Action), Czech Science Foundation [GF20-05789L] and the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement no. 873127 – InterTAU. KB is also supported by Brno Ph.D. Talent Scholarship – funded by the Brno City Municipality, Brno, Czech Republic.

References

[1] A. Bartik et al. Depicter: Intrinsic disorder and disorder function prediction server. *JMB*, 432(11):3379–3387, 2020.

[2] J. Eberhardt et al. Autodock vina 1.2.0: New docking methods, expanded force field, and python bindings. *Journal of Chemical Information and Modeling*, 2021.

[3] J. Jumper et al. Highly accurate protein structure prediction with alphafold. *Nature*, pages 583–589, 2021.

[4] L. Hongjian et al. idock: A multithreaded virtual screening tool for flexible ligand docking. *CIBCB*, pages 77–84, 2012.

[5] J. A. van Santen et al. The natural products atlas 2.0: a database of microbially-derived natural products. *Nucleic Acids Research*, 50(1):1317–1323, 2022.

[*] bhushanindu39@gmail.com
[†] Jozef.Hritz@ceitec.muni.cz
[‡] vabdna@gmail.com
[§] rajitrc@gmail.com
[¶] Krishnendu.Bera@ceitec.muni.cz

WBCB 2023, p. 12, ISBN 978-80-8147-132-2

TE-greedy-nester: a tool for reconstruction of transposable elements in plant genomes

Pavel Jedlička[1]*, Ivan Vanát[2] , Matej Lexa[1,2] , Marie Krátká[1,3] and Eduard Kejnovský[1]

[1] Department of Plant Developmental Genetics, Institute of Biophysics of the Czech Academy of Sciences, Kralovopolska 135, 61200, Brno, Czech Republic
[2] Faculty of Informatics, Masaryk University, 60200 Brno, Czech Republic
[3] National Centre for Biomolecular Research, Faculty of Science, Masaryk University, Kamenice 5, 625 00 Brno, Czech Republic

Transposable elements, including LTR retrotransposons1, are mobile genetic elements constitute remarkable portions of plant genomes and significantly contribute to genome structure, size and regulation. Because of high level of their mutual sequence similarity and numerous insertions into one another, the correct identification of full-length LTR retrotranpsons is a challenging bioinformatic task.

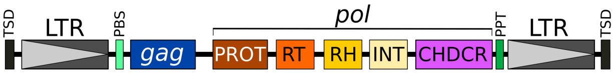

Figure 1: LTR retrotransposon scheme.

Introduction. Our software, TE-greedy-nester[1] takes advantage of greedy algorithm wich allows to mine increasingly fragmented copies of full-length LTR retrotranposons. We found this tool to be superior in computation time and full-length element recovary in highly nested regions. Using TE-greedy-nester we showed that e.g. nesting of LTR retroelements is not random[2]; nested elements often have lower LTR similarity than pre-existing ones[3]; TEs fragmentation and LTR similarity differs in (i) low-recombining Y chromosome of dioecious Silene latifolia; and (ii) in monocentric and holocentric chromosomes of closely relative species of Juncaceae family.

New features. Recent challenge is integration of tools for detection of other types of repetitive elements (e.g., tandem repeats and Miniature Inverted TEs – MITEs). Thererfore we redesigned the entire codebase to support module implementation. Thereafter, we included two other detection tools - Tandem Repeat Finder[4] and MiteFinderII[5]; and conducted a few initial testing runs on plant genomic sequences.

Acknowledgements. This work was supported by the Czech Science Foundation [grants numbers 21-00580S and 22-00364S].

References

[1] Matej Lexa, Pavel Jedlicka, Ivan Vanat, Michal Cervenansky, and Eduard Kejnovsky. TE-greedy-nester: structure-based detection of LTR retrotransposons and their nesting. *Bioinformatics*, 36(20):4991–4999, dec 2020.

[2] Pavel Jedlicka, Matej Lexa, Ivan Vanat, Roman Hobza, and Eduard Kejnovsky. Nested plant LTR retrotransposons target specific regions of other elements, while all LTR retrotransposons often target palindromes and nucleosome-occupied regions: In silico study. *Mobile DNA*, 10(1):1–14, dec 2019.

[3] Pavel Jedlicka, Matej Lexa, and Eduard Kejnovsky. What Can Long Terminal Repeats Tell Us About the Age of LTR Retrotransposons, Gene Conversion and Ectopic Recombination? *Frontiers in Plant Science*, 11:512125, may 2020.

[4] Gary Benson. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic acids research*, 27(2):573–580, jan 1999.

[5] Jialu Hu, Yan Zheng, and Xuequn Shang. MiteFinderII: a novel tool to identify miniature inverted-repeat transposable elements hidden in eukaryotic genomes. *BMC medical genomics*, 11(Suppl 5), nov 2018.

*jedlicka@ibp.cz

WBCB 2023, p. 13, ISBN 978-80-8147-132-2

Evolutionary Dynamics of LTR Retrotransposons in Holocentric Plants from Juncaceae Family

Marie Krátká[1,2]* Pavel Jedlička[1] Zdeněk Kubát[1] Eduard Kejnovský[1]

[1] Department of Plant Developmental Genetics, Institute of Biophysics of the Czech Academy of Sciences, Královopolská 135, 612 65 Brno, Czech Republic

[2] National Centre for Biomolecular Research, Faculty of Science, Masaryk University, Kamenice 5, 625 00 Brno, Czech Republic

LTR retrotransposons (Fig. 1) are mobile genetic elements which play significant role in plant genome structure, organization, and evolutionary dynamics. Holocentricity represents an unusual and distinct genome organization type (Fig. 2) which has independently arisen at least five times in Angiosperms. Expanded distribution of centromeric units in holocentrics affects the landscape of many genomic and epigenomic features [1]. Our work shows how centromere organization type affects evolutionary footprints of LTR retrotransposon activation, targeting (insertion to specific chromosomal regions or to other LTR retrotransposons), and degradation by analysing element abundance, position, insertion time, nesting, and elimination (presence of solo-LTRs) in plants from Juncaceae family — monocentric *Juncus effusus* and holocentric *Luzula sylvatica*. Analysis of this special type of chromosomes allows us to better understand the factors and mechanisms forming genomic landscape in plants.

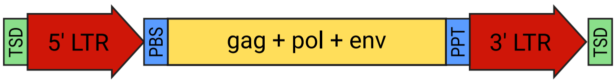

Figure 1: LTR retrotransposon structure. Element sequence consists of polypeptide genes *gag*, *pol*, and in some cases *env* flanked by direct Long Terminal Repeats (LTRs). Primer binding site (PBS) and Poly-Purine Tract (PPT) are primer sequences for reverse transcription. Insertion of the element also creates short Target Site Duplication (TSD) of the target sequence

Methods. Full-length LTR retrotransposons were identified in genome assemblies using TE-greedy-nester software [3]. Insertion time (retrotransposon age) of each element was calculated from 5'LTR and 3'LTR divergence [2]. Solo-LTRs were identified by (i) TE-greedy-nester module and (ii) annotation of LTRs by Repeatmasker as reported in Ou et al. [4].

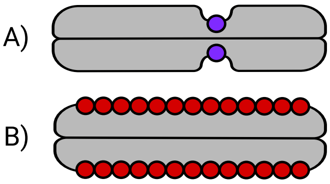

Figure 2: Centromere can be limited to a single region of the chromosome (monocentric chromosomes — A) or distributed along the entire chromosome length (holocentric chromosomes — B)

Acknowledgements. This research was supported by the Czech Science Foundation (grant 21–00580S to EK and grant 22-00364S to ZK).

References

[1] Paulo G Hofstatter, Gokilavani Thangavel, Thomas Lux, Pavel Neumann, Tihana Vondrak, Petr Novak, Meng Zhang, Lucas Costa, Marco Castellani, Alison Scott, et al. Repeat-based holocentromeres influence genome architecture and karyotype evolution. *Cell*, 185(17):3153–3168, 2022.

[2] Pavel Jedlicka, Matej Lexa, and Eduard Kejnovsky. What can long terminal repeats tell us about the age of ltr retrotransposons, gene conversion and ectopic recombination? *Frontiers in plant science*, 11:644, 2020.

[3] Matej Lexa, Radovan Lapár, Pavel Jedlička, Ivan Vanát, Michal Červeňanský, and Eduard Kejnovský. Te-nester: a recursive software tool for structure-based discovery of nested transposable elements. In *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 2776–2778. IEEE, 2018.

[4] Shujun Ou, Jinfeng Chen, and Ning Jiang. Assessing genome assembly quality using the ltr assembly index (lai). *Nucleic acids research*, 46(21):e126–e126, 2018.

*kratka@ibp.cz

Alternatives to classical HiC data analysis

Matej Lexa[1]* Hedi Hegyi[1]†

Faculty of Informatics, Masaryk University, Botanicka 63a, 600 00 Brno, Czech Republic

HiC is a variant of DNA library preparation and sequencing that can provide information about 3D arrangement patterns in chromatin.

A typical HiC experiment involves cross-linking chromatin in vivo, fragmenting it in vitro, ligating the fragments in a proximity-dependent manner. Finally a library for paired-end sequencing is created and sequenced by short read sequencing technology [3]. The typical HiC data analysis today is to preprocess HiC sequencing reads, map them to the closest reference genome, and bin mapped pairs by genomic location to obtain a contact matrix that can be visualized. Reliable tools and pipelines exist to carry out such analyses [1].

We present three alternatives to the standard HiC data processing:

- HiC Repeat Assembly Analysis (no mapping; reads are clustered/assembled instead)
- HiC Annotation Analysis (read are mapped but not binned by genomic location; binned by annotation instead),
- HiC Context Analysis (reads are mapped but not binned; instead, local flanking features associated with all contacts are identified)

In studies of eukaryotic genomes, we encountered biological questions that could be answered by analysis of HiC data, however, not by the standard procedures. Available alternatives fall into three distinct categories (above). For example, in studies of repetitive fractions of plant genomes, reads originating in repeats often cannot be reliably mapped to a unique position in a genome. An alternative can be found in clustering HiC reads together in a different manner than by mapping to specific locations. The promise of such alternative has already been shown on small data sets, using a pipeline called HiC-TE [2]. Mapping of reads to reference can be ommited or modified, while binning to overcome high noise in the data can be done according to other criteria than genomic location, or not done at all. Figure 1 shows a diagrammatic overview of these alternatives.

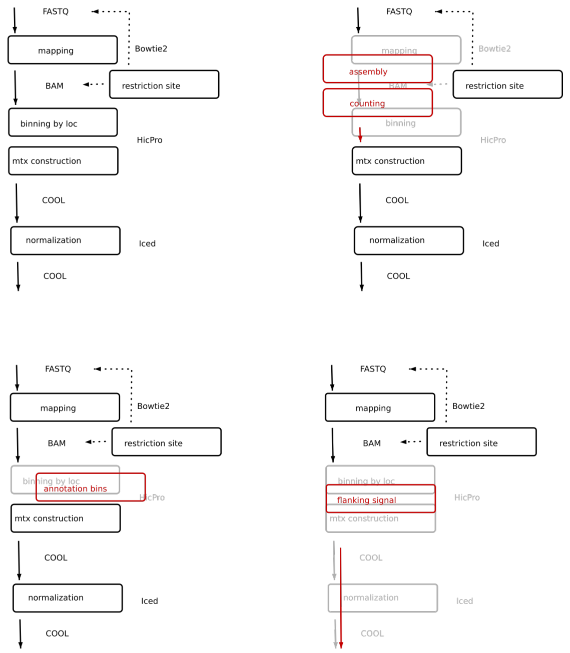

Figure 1: **Standard HiC data processing workflow (A) and the three alternative analyses (B-D).** Intermediate data formats are shown in all caps. Ommited standard steps are shown in grey.

Examples of alternative analyses will be given and discussed.

Acknowledgements. This work has been supported by the Grant Agency of the Czech Republic grant number GA21-00580S.

References

[1] M Forcato, C Nicoletti, K Pal, CM Livi, F Ferrari, and S Bicciato. Comparison of computational methods for hi-c data analysis. *Nature Methods*, 14(7):679–685, 2017.

[2] M Lexa, M Cechova, SH Nguyen, P Jedlicka, V Tokan, Z Kubat, R Hobza, and E Kejnovsky. HiC-TE: a computational pipeline for Hi-C data analysis to study the role of repeat family interactions in the genome 3D organization. *Bioinformatics*, 38(16): 4030–4032, 2022.

[3] K Pal, M Forcato, and F Ferrari. Hi-c analysis: from data generation to integration. *Biophysical Reviews*, 11:67–78, 2019.

*lexa@fi.muni.cz

†hegyihedi@gmail.com

WBCB 2023, p. 15, ISBN 978-80-8147-132-2

Using Graph Neural Networks to Detect Plasmid Contigs from an Assembly Graph

Janik Sielemann[1] Katharina Sielemann[1] Broňa Brejová[2] Tomáš Vinař[2] Cedric Chauve[3]

[1] Computational Biology, Faculty of Biology, Center for Biotechno logy (CeBiTec) & Graduate School DILS, Bielefeld Institute for Bioinformatics I nfrastructure (BIBI), Bielefeld University, 33615 Bielefeld, Germany
[2] Faculty of Mathematics, Physics and Inform atics, Comenius University in Bratislava, Slovakia
[3] Department of Mathematics, Simon Fraser University, Burnaby, Canada

We present a new method for the identification of plasmid contigs in short-read sequencing data produced from bacterial isolates. Identification of plasmids is an important and challenging problem related to antimicrobial resistance spread and other One-Health issues. Sequence assembly of short reads usually produces many short contigs, which exacerbates the difficulty of the task. However, most assemblers also provide an assembly graph showing possible connections between these contigs.

Our method uses graph neural networks to propagate information between long contigs that are easy to classify and nearby short contigs whose classification is very difficult (see Figure 1)). Most current methods are unable to classify contigs shorter than 1000 bp, while our method can perform well on contigs as short as 100 bp.

Another novel feature is the separation of plasmid and chromosome classification tasks, recognizing that some contigs are ambiguous, being parts of both types of molecules. These ambiguous contigs are an interesting subject for further study by themselves; our preliminary analysis of ambiguous contigs in our datasets suggests that the majority of them are related to transposons and phages. These mobile elements can integrate into both plasmids and chromosomes within a cell.

We have implemented our approach in an easy to use tool, called plASgraph2. Training and non-overlapping testing set included 140 and 224 short-read assemblies respectively from the ESKAPEE group of bacterial pathogenic species. In our tests, plASgraph2 outperforms Platon and PlasForest on the plasmid classification task. While plASgraph2 does not use any database or homology-based information, both Platon and PlasForest rely on such additional information, which is particularly useful for longer contigs. Yet accuracy of plASgraph2 is close to these tools even when evaluated on longer contigs.

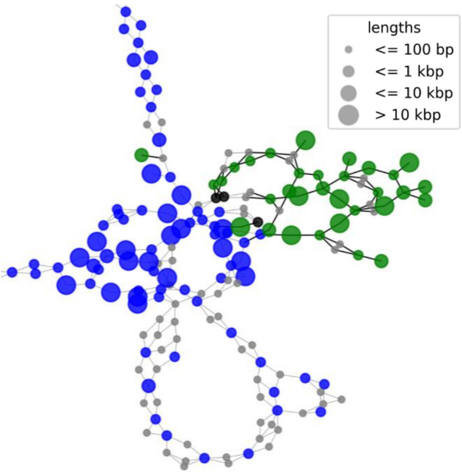

Figure 1: **Contig classification in the assembly graph of *C.freundii* isolate SAMN15148288.** Chromosomal contigs are colored in blue, plasmid contigs in green, and ambiguous contigs in black.

We have developed our tool so that training of the machine learning component is species-agnostic, and further experiments show that it can be applied successfuly to species that were not included in the training set. Thus it is an important step towards identification of previously unknown plasmids, which can be critical for diverse epidemiologic surveillance needs.

Acknowledgements. This research was supported by European Union Horizon 2020 grant No. 872539 (PANGAIA). JS and KS were supported by the Bielefeld University Graduate School Digital Infrastructure for the Life Sciences (DILS) grant. This research was also supported by grants 1/0463/20 (BB) and 1/0538/22 (TV) from the Scientific Grant Agency of the Ministry of Education, Science, Research and Sport of the Slovak Republic and Slovak Academy of Sciences (VEGA), grant APVV-22-0144 from the Slovak Research and Development Agency (BB, TV), and Discovery Grant RGPIN/03986-2017 from the Natural Sciences and Engineering Research Council of Canada (CC). This research was enabled in part by computational infrastructure support provided by Digital Research Alliance of Canada (`https://alliancecan.ca`).