M U N I

Annex No. 10 to the MU Directive on Habilitation Procedures and Professor Appointment Procedures

HABILITATION THESIS REVIEWER'S REPORT

Masaryk University	
Applicant	Matej Lexa, Doctor of Philosophy
Habilitation thesis	Algorithmic approaches to biological sequence analysis generate new tools for the study of genome structure and function
Reviewer	**Prof. Peter F. Stadler**
Reviewer's home unit, institution	Institut for Informatik, University Leipzig, Germany

The habilitation thesis submitted by Matej Lexa with the title "Algorithmic approaches to biological sequence analysis generate new tools for the study of genome structure and function" covers six research papers as a representative selection of the applicant's scientific achievements over last two decades. The material included shows that the Matej Lexa has successfully addressed key issues in modern bioinformatics. His work focusses on the analysis of DNA sequences with the aim of identifying functional subsequences, and on the design of efficient algorithms for such tasks in conjunction with the implementation of practically usable tools. In the following paragraphs let me briefly consider each of the six publications included in the thesis.

[1] Virtual PCR, Bioinformatics 2001. While primer design tools where available, they did not consider what is now commonly called off-target effects. In the paper, Matej Lexa and co-authors describe the first tool to predict PCR products for an entire genome as input. VPCR explicitly simulated the PCR cycles as a system of ODEs, based candidate regions identified by blast. The approach has been widely used by others in follow-up work, making it a noticeable contribution to the field.

[2] A k-mer based algorithm for pattern matching, Bioinformatics 2003, in essence replaces the blast-based candidate search by hashing k-mers. Following ideas from blast, PRIMEX uses pairs of approximate k-mer matches. The method improved the candidate's own Virtual PCR approach. It was also used successfully to delineate protein domains. Like the previous paper, it was a significant step forward at the time, albeit its use has largely been superseeded by modern short-read aligners, which use more efficient index data structures.

[3] A triple-stranded DNA pattern matching, Bioinformatics 2011. The work is based on the realization that triple-helical regions can be understood as pairs of pairwise alignments, with Watson-Crick and Hogsteen pairs taking the role of matches. Since in DNA the Watson-Crick paired regions are perfectly complementary, the problem reduces a palindrom-search closely related to the computation of circular matchings, i.e. RNA folding, for each of the eight basis topologies of H-DNA. The Triplex software is still in use.

[4] A branch and bound G-quadruplex matching algorithm, Bioinformatics 2017. The analysis of G-quadruplexes in both DNA and RNA were a "hot topic" at the time PQSfinder was published. The tool implements a branch and bound algorithm that exhaustively lists all plausible Quadruplex structures including non-canonical variants with bulges. The approach is surprisingly accurate given its simplicity, and is still a well-used tool.

[5] A greedy algorithm for detection of nested structures in genomic DNA sequences, Bioinformatics 2020, is concerned with disentangling the nesting of transposons that are inserted into transposons. The straightforward greedy approach starts from identifying full-length uninterrupted elements, identifies and removes them from the target sequence, and continues on the remaining sequence. Much work went into tuning parameters and the adjustment of cut-offs as the algorithm progresses.

[6] Nextflow pipeline for Genome annotation workflow, Bioinformatics 2022, is concerned with the use of spatial HiC data for characterization of repetitive sequences. The HiC-TE pipeline collects statistics on the interaction of repeats at the level of repeat families. A key issue is the normalization of the count data to account for various biases inherent in HiC data. The work helps to analyse the organization of repetitive material, a topic that is still under-studied in (comparative) genomics.

Overall, the six contributions show-cased in the thesis outline a coherent research agenda that aims at the efficient use of string-algorithms in genomics. While the algorithms themselves are neither particularly surprising nor difficult, they are well-tailored towards the intended applications. Matej Lexa has made contributions that are scientifically sound, timely, useful to the bioinformatics community, and rather well cited.

The habilitation thesis is significantly different from those I have seen as a reviewer from other countries, with a very short general introduction and almost autobiographic snippets accompanying the individual chapters. I assume that this format follows the customs at Masaryk University. In any case, the scientific content of Matej Lexa's work is, in my opinion, clearly sufficient to fulfil the requirements expected of a habilitation thesis in the research area of bioinformatics.

Reviewer's questions for the habilitation thesis defence (number of questions up to the reviewer)

(1) In the light of the recent advances in short-read aligners, would you use a different algorithmic approach, e.g. suffix arrays or a similar index structure, to address off-target PCR products?

(2) Triple helices also appear in RNA, where the situation is more difficult than in DNA since the Watson-Crick duplex does not need to be perfectly complementary but may include GU pairs and possibly even small bulges. To what extent is your approach still applicable, or would you consider a very different algorithmic approach?

(3) How realistic are the constraints on the bulge sizes in the quadruplex-finding algorithms? Is the approach still feasible if longer bulges have to be allowed? How non-local is the scoring? Does it make sense to think about a dynamic programming algorithm to handle overlapping candidate quadruplexes?

(4) In what sense is the greedy approach of TE-greedy-nester not optimal? What would an exact algorithm look like? It wound be interesting to consider the nested accumulation of TEs in a comparative genomic setting. Do you have any ideas how to distinguish TE insertions before and after a speciation event that separates two sequenced genomes?

(5) Is the large signal along the diagonal still a reflection of the proximity bias, or is this an indication that TEs often proliferate by tandem duplications? Is there a way to distinguish from the HiC data whether genomically proximal or distal TEs are in contact?

Conclusion

The habilitation thesis entitled “Algorithmic approaches to biological sequence analysis generate new tools for the study of genome structure and function” by Matej Lexa, Doctor of Philosophy, fulfils the requirements expected of a habilitation thesis in the field of Informatics.

Date: 05.04.2023