FP7-PEOPLE-IRG 224285 ALGGENOMES
Algorithms for Analysis of Genes and Genomes

Tomas Vinar, PhD.
Faculty of Mathematics, Physics, and Informatics, Comenius University,
Mlynska Dolina, 842 48 Bratislava, Slovakia
E-mail: vinar@fmph.uniba.sk
Project website: `http://compbio.fmph.uniba.sk/`

Publishable Summary January 15, 2009 – January 14, 2013

ALGGENOMES is a Marie Curie project supporting reintegration of Dr. T. Vinar at the Faculty of Mathematics, Physics and Informatics of Comenius University in Bratislava, Slovakia. Dr. Vinar has spent almost ten years of his research career in Canada and the United States after which he decided to return to Slovakia in order to start a bioinformatics research and education program there. The main goals of the project, as outlined in the grant agreement, are as follows:

A: **Development of algorithms for bioinformatics.** Design of new algorithms and probabilistic models for a variety of bioinformatics problems in sequence analysis and gene evolution, their implementation, and application of the resulting tools to the analysis of real biological datasets.

B: **Analysis of yeast mitochondrial genomes.** A collaboration with the research group of Prof. Nosek at the Faculty of Natural Science to study evolution of mitochondrial genomes of pathogenic yeasts, with the focus on rearrangements.

C: **Supporting activities.** Setup and development of a computational environment necessary for bioinformatics research, recruitment and supervision of students, and teaching activities supporting the research in this proposal.

We have developed new algorithms for comparative genomic analysis of complex duplicated regions (goal A): an algorithm for reconstruction of evolutionary histories of gene clusters (Vinar et al., 2010), an artificial simulation framework, and an algorithm for automated segmentation of gene clusters (Brejova et al., 2011a). We have applied these algorithms to analyze the evolutionary history of the alpha defensin gene cluster in the primate genomes (Orangutan Genome Sequencing Consortium, 2011) and we have also investigated other theoretical and algorithmic problems stemming out of this research (Brejova et al., 2011b; Kovac et al., 2012). Together with our collaborators from Penn State University and National Human Genome Research Institute, we have become members of an ongoing collaboration for sequencing and analysis of biomedically important complex gene clusters. We are working on improved algorithms for gene cluster analysis through more efficient MCMC sampling, and on making our prototype software tools available to a wider community.

In collaboration with Dr. Luptak at the University of California at Irvine, we have developed a new sequence analysis algorithm and a software for RNA motif search (Jimenez et al., 2012) that is currently being applied in biochemical research on ribozymes. We have used our experience in RNA motif search to develop a similar framework for contact-rich protein domain search (Macko et al., 2013). We have also studied theoretical and practical problems relevant to annotation of alternative splicing (Kovac et al., 2009) and gene finding in novel genomes (Brejova et al., 2009). Finally, we have continued developing software for identification of gene orthologs and methodology for studying positive selection, which we have applied in several international projects (Panda Genome Sequencing and Analysis Consortium, 2010; Orangutan Genome Sequencing Consortium, 2011; The Western Painted Turtle Genome Consortium, 2013; The Marmoset Genome Sequencing and Analysis Consortium, 2013).

In collaboration with the Laboratory of Comparative and Functional Genomics of Eukaryotic Organelles (prof. Nosek, goal B) and with Dr. Brejova at the Department of Computer Science, we have analyzed eight newly sequenced mitochondrial genomes of pathogenic yeasts, with focus on their phylogeny and rearrangement history (Valach et al., 2011). To this end, we have developed a novel algorithm and software for analysis of rearrangement histories based on double-cut-and-join rearrangement model (Kovac et al., 2011a) and we also studied several theoretical problems in this area (Kovac et al., 2010, 2011b; Jahn et al., 2012).

Within the supporting activities (goal C), we have succesfuly established a computational biology research group at the Faculty of Mathematics, Physics, and Informatics that comprises two principal investigators (Dr. Vinar and Dr. Brejova) and 16 students at all levels of studies (bachelor's, master's, doctolar). Dr. Vinar currently supervises research projects of three doctoral students (Jakub Kovac, Martin Macko, Martin Kravec), and five master students; two bachelor and five masters theses have been completed within the scope of this project so far. The research group maintains stable research collaborations with scientists from Austria, China, Germany, Canada, and the Unites States. With contribution from a separate grant to Dr. Brejova, we have built a research computing cluster that supports the activities of the research group. To support our research activities, we maintain a weekly seminar on recent topics in computational biology, as well as regular lab meetings.

In collaboration with Dr. Brejova (Dept. of Computer Science), prof. Nosek, and prof. Tomaska (Faculty of Natural Sciences), we have developed two courses covering the area of computational biology ("Methods in Bioinformatics" and "Genomics"), and we have started preparations for establishing bioinformatics degree program. We organize common seminars and summer schools. These educational activities (goal C) will ensure sustainability of research in computational biology at our institution.

References

Brejova, B., Burger, M., and Vinar, T. (2011a). Automated Segmentation of DNA Sequences with Complex Evolutionary Histories. In Przytycka, T. M. and Sagot, M.-F., editors, *Algorithms in Bioinformatics, 11th International Workshop (WABI)*, volume 6833 of *Lecture Notes in Computer Science*, pages 1–13, Saarbrcken, Germany. Springer.

Brejova, B., Landau, G. M., and Vinar, T. (2011b). Fast Computation of a String Duplication History under No-Breakpoint-Reuse. In Grossi, R., Sebastiani, F., and Silvestri, F., editors, *String Processing and Information Retrieval (SPIRE)*, volume 7024 of *Lecture Notes in Computer Science*, pages 144–155, Pisa, Italy. Springer.

Brejova, B., Vinar, T., Chen, Y., Wang, S., Zhao, G., Brown, D. G., Li, M., and Zhou, Y. (2009). Finding genes in Schistosoma japonicum: annotating novel genomes with help of extrinsic evidence. *Nucleic Acids Research*, 37(7):e52.

Jahn, K., Zheng, C., Kovac, J., and Sankoff, D. (2012). A consolidation algorithm for genomes fractionated after higher orderpolyploidization. *BMC bioinformatics*, 13(Suppl 19):S8.

Jimenez, R. M., Rampasek, L., Brejova, B., Vinar, T., and Luptak, A. (2012). Discovery of RNA Motifs Using a Computational Pipeline that Allows Insertions in Paired Regions and Filtering of Candidate Sequences . In Hartig, J. S., editor, *Ribozymes: Methods and Protocols*, volume 848 of *Methods in Molecular Biology*, chapter 10, pages 145–158. Springer.

Kovac, J., Braga, M. D. V., and Stoye, J. (2010). The Problem of Chromosome Reincorporation in DCJ Sorting and Halving. In Tannier, E., editor, *Comparative Genomics - International Workshop, RECOMB-CG*, volume 6398 of *Lecture Notes in Computer Science*, pages 13–24, Ottawa, Canada. Springer.

Kovac, J., Brejova, B., and Vinar, T. (2011a). A Practical Algorithm for Ancestral Rearrangement Reconstruction. In Przytycka, T. M. and Sagot, M.-F., editors, *Algorithms in Bioinformatics, 11th International Workshop (WABI)*, volume 6833 of *Lecture Notes in Computer Science*, pages 163–174, Saarbrcken, Germany. Springer.

Kovac, J., Vinar, T., and Brejova, B. (2009). Predicting gene structures from multiple RT-PCR tests. In *Algorithms in Bioinformatics (WABI)*, volume 5724 of *Lecture Notes in Bioinformatics*, pages 181–193. Springer.

Kovac, J., Warren, R., Braga, M. D. V., and Stoye, J. (2011b). Restricted DCJ Model: Rearrangement Problems with Chromosome Reincorporation. *Journal of Computational Biology*, 18(9):1231–1231.

Kovac, P., Brejova, B., and Vinar, T. (2012). Aligning Sequences with Repetitive Motifs. In *Information Technologies - Applications and Theory (ITAT)*, pages 41–48. Best paper award.

Macko, M., Kralik, M., Brejova, B., and Vinar, T. (2013). OB-Fold Recognition Combining Sequence and Structural Motifs. Submitted, under review.

Orangutan Genome Sequencing Consortium (2011). Comparative and demographic analysis of orang-utan genomes. *Nature*, 469(7331):529–533.

Panda Genome Sequencing and Analysis Consortium (2010). The sequence and de novo assembly of the giant panda genome. *Nature*, 463(7279):269–392.

The Marmoset Genome Sequencing and Analysis Consortium (2013). The Genome of the Common Marmoset: A Comparative Analysis of an Extraordinary South American Primate. Submitted, under review.

The Western Painted Turtle Genome Consortium (2013). The Western Painted Turtle Genome: The Evolution of Extreme Physiological Adaptations in a Slowly Evolving Lineage. Submitted, under review.

Valach, M., Farkas, Z., Fricova, D., Kovac, J., Brejova, B., Vinar, T., Pfeiffer, I., Kucsera, J., Tomaska, L., Lang, B. F., and Nosek, J. (2011). Evolution of linear chromosomes and multipartite genomes in yeast mitochondria. *Nucleic Acids Research*, 39(10):4202–4219.

Vinar, T., Brejova, B., Song, G., and Siepel, A. C. (2010). Reconstructing histories of complex gene clusters on a phylogeny. *Journal of Computational Biology*, 17(9):1267–1279.