Distance and Similarity Measures in Comparative Genomics

Diego P. Rubert[1], **Jens Stoye**[2], **Fábio H. V. Martinez**[1]

[1]Faculdade de Computação – UFMS, Campo Grande, MS – Brazil

[2]Faculty of Technology – Bielefeld University, Bielefeld – Germany

{diego,fhvm}@facom.ufms.br, jens.stoye@uni-bielefeld.de

Abstract. *Research in comparative genomics supports the investigation of important questions in molecular biology, genetics and biomedicine. A central question in this field is the elucidation of similarities and differences between genomes by means of different measures. This summary describes the main contributions, originality and impact possibilities of the thesis entitled "Distance and Similarity Measures in Comparative Genomics", by Diego P. Rubert.*

1. Introduction

Comparative genomics is a field of biological research in which genomic features, such as DNA sequence, genes, gene order, regulatory sequences or other structural aspects, are evaluated to compare different species. To this end, whole or parts of genomes are compared to find biological similarities and differences as well as evolutionary relationships between organisms. Based on these findings, genome and molecular evolution can be inferred and this may in turn be put in the context of, for instance, phenotypic evolution, phylogenetic evaluation, population genetics or ancestral genome reconstruction.

Genome rearrangements—large-scale mutations responsible for complex changes and structural variations—are subject of extensive studies in comparative genomics. A classical problem in comparative genomics is, given two genomes, to compute the rearrangement *distance* between them, i.e., finding the minimum number of rearrangement operations that transform one genome into the other, or the genomic *similarity* between them, i.e., finding maximally similar regions given some rearrangement operation. The study of these problems supports the investigation of important questions in a number of other fields, such as molecular biology, genetics, biomedicine and paleogenomics.

A number of known rearrangement operations are known. In 2005 a novel powerful model that can represent most of large-scale rearrangements operations (see Figure 1) was proposed by Yancopoulos, Attie and Friedberg [Yancopoulos et al. 2005]: the *double cut and join* (DCJ). It consists of cutting a genome in two distinct positions (possibly in two distinct chromosomes) and joining the four resultant open ends in a different way.

The purpose of the doctoral thesis was to investigate distance and similarity measures for genomes under DCJ events and applications of such measures. The most basic scenario in comparing genomes is where each gene occurs only once in each of the genomes. In this scenario, several measures have been studied, many of which can be efficiently computed. However, this model does not match strictly what is found in nature, where several copies of the same gene (or orthologous genes) occur in the same genome due to duplications. When orthologous genes occur, *gene families* are defined, which are gene sets with similar biochemical functions. In order to address the occurrence of

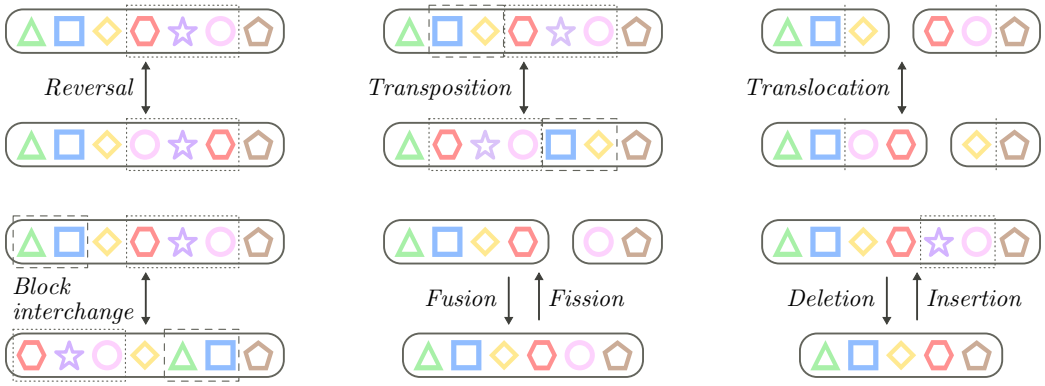

Figure 1. Examples of rearrangement events. Symbols are genes, solid lines group genes in chromosomes and dashed or dotted lines represent regions or positions affected by a rearrangement event. One DCJ can mimic a reversal, a fusion, a fission, or a translocation, whereas two DCJs are required to model a block interchange or a transposition.

multiple genes from the same family in genomes, the *family-based* approaches have been proposed, establishing measures usually hard to compute.

Although family information can be obtained by accessing public databases or by direct computing, data can be incorrect, and inaccurate families could be providing support to erroneous assumptions of homology between segments. Thus, it is not always possible to classify each gene unambiguously into a single family, and an alternative to the family-based setting was proposed recently [Doerr et al. 2012, Braga et al. 2013, Martinez et al. 2015]. It consists of studying genome rearrangements without prior family assignment, by directly accessing the pairwise similarities between genes of the compared genomes. This approach is said to be *family-free* (FF). Such problems in general are at least as difficult as those based on gene families.

Given this context, the main outcomes of the thesis can be summarized bellow. In the following sections, each one of the main outcomes are further discussed. A last section enumerates the main byproducts of the doctoral thesis.

1. A linear time approximation algorithm with approximation ratio $O(k)$ for a restrict case of the family-based DCJ distance problem. The general case of this problem is NP-hard and there already exists an ILP exact algorithm for solving it. Experiments show that the approximation algorithm is very competitive both in efficiency and in quality of the solutions with respect to the ILP algorithm.
2. The APX-hardness proof for the family-free DCJ similarity measure, an exact ILP algorithm and four combinatorial heuristics to solve the problem. This problem was previously known to be NP-hard. Experiments show that the proposed heuristics are fast and return good results compared to the proposed ILP algorithm.
3. A novel local similarity measure based on DCJ operations, the local DCJ similarity. Analogous to local sequence alignment, the local DCJ similarity scores local regions in compared genomes with high levels of structural similarity. Such a local measure is often convenient when comparing highly dissimilar genomes containing some or many conserved regions. We show its usefulness by modifying a popular ancestral genome reconstruction pipeline, performing the ancestral

reconstruction for an eudicots dataset, and obtaining improved results compared to those presented in a recent publication using the original pipeline.

All main outcomes have potential to directly improve the ways and how efficiently genomes are compared, allowing researchers to not only trace out the evolutionary relationship between organisms but also differences and similarities within and between species. For instance, comparison of the fruit fly and human genomes reveals that two-thirds of human genes known to be involved in cancer have counterparts in the fruit fly. Besides the application for human health, improving the ways genomes are compared may widely benefit ecological studies. As genome sequencing grows easier and less expensive, genome comparison allows global views on genome evolution and helps in deciphering the hidden information in genome design, function and evolution, impacting applications in agriculture, biotechnology, and zoology as a tool to tease apart the often-subtle differences among animal and plant species. Such efforts also contribute to the rearrangement of our understanding of some branches of the evolutionary "tree of life", and helps finding new strategies for conserving rare and endangered species. Besides leading to a better understanding of how species evolved, comparing the genomes of different species one can determine the function of novel un-annotated genes and noncoding regions in genomes.

2. Family-based DCJ distance

We have presented an approximation algorithm for computing the family-based DCJ distance between two balanced[1] unichromosomal genomes with at most k copies of each gene [Rubert et al. 2016, Rubert et al. 2017a], an NP-hard problem [Shao et al. 2015]. The main step of the approximation algorithm relies on an approximation for an related problem, the *Breakpoint Distance* (BD), and on decomposing a special bipartite graph called *adjacency graph* (AG) into cycles.

Let k be the maximum number of occurrences (or copies) of any gene in the input genomes. For general values of k, BD has an $O(k)$-approximation [Jiang et al. 2010]. The latter result is based on a linear time approximation algorithm for a restricted version of the Minimum Common String Partition (k-MCSP) problem with approximation ratio $O(k)$ [Kolman and Waleń 2007].

We have shown that the algorithm we have developed, named CONSISTENT-DECOMPOSITION, also has an approximation ratio $O(k)$:

Theorem 3.3 *Let A and B be balanced unichromosomal linear genomes with at most k copies of each gene. Let $(\mathcal{A}, \mathcal{B})$ be a common string partition with approximation ratio $O(k)$ for k-MCSP$(\widehat{A}, \widehat{B})$. A consistent decomposition D of $AG(A, B)$, containing cycles of length 2 reflecting preserved adjacencies in $(\mathcal{A}, \mathcal{B})$, is an $O(k)$-approximation for the* DCJ-DISTANCE *problem.*

Besides, it works properly for linear and circular genomes and runs in linear time:

Theorem 3.6 *Given balanced unichromosomal linear genomes A and B such that $|A| = |B| = n$ and at most k copies of each gene, the running time of the algorithm* CONSISTENT-DECOMPOSITION *is linear in the size of the genomes, i.e., $O(n)$.*

[1] Each gene has the same number of copies or occurrences in each genome.

Experiments on simulated data sets show that the approximation algorithm is very competitive both in efficiency and in quality of the solutions when compared to an exact ILP algorithm. In addition, the approximation algorithm runs always in a fraction of a second and therefore could be used to provide an initial lower bound to the ILP solver.

3. Family-free DCJ similarity

We have also studied the problem of computing the overall similarity of two given linear or circular multichromosomal genomes in a family-free setting under the DCJ model. This problem is called FFDCJ-SIMILARITY. The complexity of computing the FFDCJ-SIMILARITY was proven to be NP-hard [Martinez et al. 2015], while the counterpart problem of computing the family-free DCJ distance was already proven to be APX-hard.

We have demonstrated FFDCJ-SIMILARITY is APX-hard and an inapproximability result by a strict (approximation-preserving) reduction from MAX-2SAT to FFDCJ-SIMILARITY [Rubert et al. 2017b, Rubert et al. 2018]:

Theorem 4.4 FFDCJ-SIMILARITY *is APX-hard.*

Corollary 4.6 FFDCJ-SIMILARITY *cannot be approximated with approximation ratio better than* $22/21 = 1.0476\ldots$*, unless* $P = NP$*.*

We then present an exact ILP algorithm to solve it and four combinatorial heuristics, with computational experiments comparing their results for datasets simulated by a framework for genome evolution.

While the ILP is fast and accurate for smaller instances, it cannot solve larger instances due to the number of restrictions, which is cubic in the size of the input genomes. On the other hand, the heuristics obtained good results for a number of instances, some of them even better than the results returned by the ILP solver after reaching the time limit. Moreover, the heuristics presented fast running times for realistic-size genomes. The ILP could benefit greatly from heuristics by using their outputs as initial lower bounds.

4. Family-based local DCJ similarity

Lastly, we have introduced the concept of (family-based) local DCJ similarity, a local genome rearrangement measure analog to local sequence alignment. Then, we have presented ANGORA, an improved workflow for ancestral genome reconstruction from highly diverged genomes such as those of plants, showing how the local DCJ similarity allowed reconstructed ancestral regions with higher resolution [Rubert et al. 2020]. Such a workflow relies on an established workflow in the reconstruction of ancestral plants [Salse 2016]. However, it is important to stress that the enhancement allowed by the local DCJ similarity is only one among other improvements of our workflow.

More specifically, first, instead of using annotated genes, we identify genomic markers and use them as building blocks of the ancestral sequence, allowing us to reconstruct both intra- and intergenic blocks of DNA. This enables us to reconstruct the ancestral genome from hundreds of thousands of markers rather than the tens of thousands of annotated genes. Second, instead of using CloseUp, a statistical method for discovering syntenic blocks in pairs of genomic sequences, we use Gecko3 [Winter et al. 2016], which computes exact solutions under a principled definition of synteny [Jahn 2011] in

multiple sequences. Third, based on the local DCJ similarity, we score syntenic blocks and refine the family assignment of their contained markers.

Recently, Badouin and colleagues reconstructed the ancestral genome of eudicots, a major sub-clade of flowering plants, from the gene annotations of grape, coffee, artichoke, lettuce and sunflower and arrived at an ancestral genome comprising $6{,}525$ genes [Badouin et al. 2017]. With the enhanced workflow at hand, we have reconstructed the same ancestral genome. Our reconstructed genome is highly detailed, yet its layout agrees well with the reference reconstruction, reported by Badouin and colleagues, which was made using the original workflow.

Our improvements led to a reconstruction of the ancestral eudicot genome that is composed of $32{,}788$ markers distributed across $3{,}153$ contiguous ancestral regions (CARs), which are regions supposed to be in the ancestral genome. Remarkably, the layout of our ancestral genome differs on average only in 3.2% from that Badouin *et al.* Our method is also applicable to gene-based reconstruction, where it increased the genome content of the eudicot ancestor to $6{,}961$ reconstructed genes while differing on average only in 4.6% from Badouin *et al.*'s reconstruction. In other words, using local genome rearrangement, not only the marker-based, but also the gene-based reconstruction of the ancestor exhibited increased genome content, evidencing the power of this novel concept.

5. Byproducts of the doctoral thesis

Submissions accepted in international conferences and journals, and a list of produced software are listed in the following.

Submissions accepted for presentation in international conferences:

1. A linear time approximation algorithm for the DCJ distance for genomes with bounded number of duplicates. In *Proc. of the 16th Intl. Workshop on Algorithms in Bioinformatics (WABI 2016)*, 2016. [Rubert et al. 2016]
2. Algorithms for computing the family-free genomic similarity under DCJ. In *Proc. of the 15th RECOMB Comparative Genomics Satellite Intl. Workshop (RECOMB-CG 2017)*, 2017. [Rubert et al. 2017b]
3. Analysis of local genome rearrangement improves resolution of ancestral genomic maps in plants. In *17th RECOMB Comparative Genomics Satellite Intl. Workshop RECOMB-CG 2019)*. (published in BMC Genomics, see bellow)

RECOMB-CG and WABI are long-established and prestigious conference series in bioinformatics. Some of the most prominent names in the field are involved and regularly publish their best papers there.

Publications in journals:

1. Approximating the DCJ distance of balanced genomes in linear time. *Algorithms for Molecular Biology*, 2017. [Rubert et al. 2017a]
2. Computing the family-free DCJ similarity. *BMC Bioinformatics*, 2018. [Rubert et al. 2018]
3. Analysis of local genome rearrangement improves resolution of ancestral genomic maps in plants. *BMC Genomics*, 2020. [Rubert et al. 2020]

Algorithms for Molecular Biology, BMC Bioinformatics and BMC Genomics are among the most traditional and top indexed journals in bioinformatics.

Software:

1. `k-dcj-approx-dup` – A linear time approximation for the DCJ distance for balanced genomes with at most k duplicates;
2. `ffdcj-sim` – An exact ILP algorithm and heuristics for the family-free DCJ similarity;
3. `GEESE` – An efficient parallel implementation for GEnomE SEgmentation;
4. `Gecko3-DCJ` – Finds approximate gene clusters and analyzes structural similarity by the local DCJ similarity;
5. `ANGORA` – ANcestral reconstruction by local GenOme Rearrangement Analysis.

The software above are publicly available under the terms of the GNU GPL.

References

Badouin, H. et al. (2017). The sunflower genome provides insights into oil metabolism, flowering and Asterid evolution. *Nature*, 546(7656):148–52.

Braga, M. D. V., Chauve, C., Dörr, D., Jahn, K., Stoye, J., Thévenin, A., and Wittler, R. (2013). The potential of family-free genome comparison. In Chauve, C., El-Mabrouk, N., and Tannier, E., editors, *Models and Algorithms for Genome Evolution*, chapter 13.

Doerr, D., Thévenin, A., and Stoye, J. (2012). Gene family assignment-free comparative genomics. *BMC Bioinformatics*, 13(Suppl 19):S3.

Jahn, K. (2011). Efficient computation of approximate gene clusters based on reference occurrences. *Journal of Computational Biology*, 18(9):1255–1274.

Jiang, H., Zheng, C., Sankoff, D., and Zhu, B. (2010). Scaffold filling under the breakpoint distance. In *Proc. of the 8th RECOMB Comparative Genomics Satellite International Workshop (RECOMB-CG 2010)*, volume 6398 of *LNB*, pages 83–92.

Kolman, P. and Waleń, T. (2007). Reversal distance for strings with duplicates: Linear time approximation using hitting set. *The Electronic Journal of Combinatorics*, 14.

Martinez, F. V., Feijão, P., Braga, M. D. V., and Stoye, J. (2015). On the family-free DCJ distance and similarity. *Algorithms for Molecular Biology*, 10:13.

Rubert, D. P., Feijão, P., Braga, M. D. V., Stoye, J., and Martinez, F. H. V. (2017a). Approximating the DCJ distance of balanced genomes in linear time. *Algorithms for Molecular Biology*, 12(1):3.

Rubert, D. P., Feijão, P., Braga, M. D. V., Stoye, J., and Martinez, F. V. (2016). A linear time approximation algorithm for the DCJ distance for genomes with bounded number of duplicates. In *Proc. of the 16th International Workshop on Algorithms in Bioinformatics (WABI 2016)*, pages 293–306.

Rubert, D. P., Hoshino, E. A., Braga, M. D. V., Stoye, J., and Martinez, F. V. (2018). Computing the family-free DCJ similarity. *BMC Bioinformatics*, 19(6):152.

Rubert, D. P., Martinez, F. V., Stoye, J., and Doerr, D. (2020). Analysis of local genome rearrangement improves resolution of ancestral genomic maps in plants. *BMC Genomics*, 21(S2):273. (Proc. of the 17th RECOMB Comparative Genomics Satellite International Workshop RECOMB-CG 2019).

Rubert, D. P., Medeiros, G. L., Hoshino, E. A., Braga, M. D. V., Stoye, J., and Martinez, F. V. (2017b). Algorithms for computing the family-free genomic similarity under DCJ. In *Proc. of the 15th RECOMB Comparative Genomics Satellite International Workshop (RECOMB-CG 2017)*, pages 76–100.

Salse, J. (2016). Ancestors of modern plant crops. *Current Opinion in Plant Biology*, 30:134–42.

Shao, M., Lin, Y., and Moret, B. (2015). An exact algorithm to compute the double-cut-and-join distance for genomes with duplicate genes. *Journal of Computational Biology*, 22(5):425–435.

Winter, S., Jahn, K., Wehner, S., Kuchenbecker, L., Marz, M., Stoye, J., and Böcker, S. (2016). Finding approximate gene clusters with GECKO 3. *Nucleic Acids Res.*, 44.

Yancopoulos, S., Attie, O., and Friedberg, R. (2005). Efficient sorting of genomic permutations by translocation, inversion and block interchanges. *Bioinformatics*, 21(16):3340–3346.