Algorithms for Sorting by Reversals or Transpositions, with Application to Genome Rearrangement

Gustavo Rodrigues Galvão (Author)[1] **and Zanoni Dias (Supervisor)**[1]

[1]Institute of Computing
University of Campinas (Unicamp) – Campinas, SP – Brazil

{ggalvao,zanoni}@ic.unicamp.br

Abstract. *The problem of finding the minimum sequence of rearrangements that transforms one genome into another is a well-studied problem that finds application in comparative genomics. Representing genomes as permutations, in which genes appear as elements, that problem can be reduced to the combinatorial problem of sorting a permutation using a minimum number of rearrangements. Such combinatorial problem varies according to the types of rearrangements considered. The PhD thesis summarized in this paper presents exact, approximation, and heuristic algorithms for solving variants of the permutation sorting problem involving two types of rearrangements: reversals and transpositions.*

1. Introduction

One of the challenges of modern science is to understand how species evolve. As evolution can be viewed as a branching process, whereby new species arise from changes occurring in living organisms, the study of the evolutionary history of a group of species is commonly made by analyzing trees whose nodes represent species and edges represent evolutionary relationships. Since these relationships are referred to as phylogeny, such trees are called phylogenetic trees.

Phylogenies can be inferred from different kinds of data, from geographic and ecological, through behavioral, morphological, and metabolic, to molecular data, such as DNA. Molecular data have the advantage of being exact and reproducible, at least within experimental error, not to mention fairly easy to obtain [Gascuel 2005, Chapter 12]. Distance-based methods form one of the three large groups of methods to infer phylogenetic trees from molecular or sequence data [Lemey et al. 2009, Chapter 5]. Such methods proceed in two steps. First, the evolutionary distance is computed for every sequence pair and this information is stored in a matrix of pairwise distances. Then, a phylogenetic tree is constructed from this matrix using a specific algorithm, such as *Neighbor-Joining* [Saitou and Nei 1987]. Note that, in order to complete the first step, we need some method to estimate the evolutionary distance between a sequence pair. Assuming the sequence data correspond to complete genomes, we can resort to the genome rearrangement approach [Fertin et al. 2009] in order to estimate the evolutionary distance.

Using the genome rearrangement approach, one estimates the evolutionary distance between two genomes by finding the rearrangement distance between them, which is the length of the shortest sequence of rearrangement operations that transforms one genome into the other. Assuming genomes consist of a single chromosome, share the same set of genes, and contain no duplicated genes, we can represent them as permutations of integers, where each integer corresponds to a gene. If, besides the order, the

orientation of the genes is also considered, then each integer has a sign, $+$ or $-$, and the permutation is called a signed permutation. Similarly, we also refer to a permutation as an unsigned permutation when its elements do not have signs. Moreover, if the genomes are circular, then the permutations are also circular; otherwise, they are linear.

By representing genomes as permutations, the problem of finding the shortest sequence of operations that transforms one genome into another can be reduced to the combinatorial problem of computing the minimum number of operations necessary to transform one permutation into another. By algebraic properties of permutations, this problem can be equivalently stated as the problem of computing the minimum number of operations necessary to transform one permutation into the identity permutation $(+1\ +2\ \ldots\ +n)$. This problem is commonly referred to as the permutation sorting problem or as the rearrangement sorting problem.

Depending on the operations allowed to sort a permutation, we have a different variant of the permutation sorting problem. The PhD thesis defended by the author [Galvão 2015] focus on solving variants that take into account two types of operations: reversals and transpositions. A reversal is responsible for reversing the order and flipping the signs of a sequence of elements within a permutation, while a transposition is responsible for switching the location of two contiguous portions of a permutation. They are the most often considered operations for rearrangement-based phylogenetic reconstruction.

2. Contributions and Organization

The thesis consists of a collection of 5 articles that were published in peer-reviewed journals and conference proceedings in the course of the PhD. Specifically, the thesis contains 7 chapters: an introduction followed by 5 chapters (each one corresponding to one article) and a conclusion. The following paragraphs summarize the contents of the intermediary chapters, highlighting the main contributions.

Chapter 2 corresponds to an article [Galvão and Dias 2014b] published in *ACM Journal of Experimental Algorithmics*. In this chapter, we present a tool, called GRAAu, to audit algorithms for permutation sorting problems. The audit consists in comparing, for all permutations of up to a given size, the distance outputted by a given algorithm with the related rearrangement distance, and then producing statistics that can be used to analyze the performance of this algorithm. We also present tightness results for some approximation algorithms regarding two variants of the permutation sorting problem: the problem of sorting by prefix reversals and the problem of sorting by prefix transpositions. Part of the results presented in this chapter were also presented in the Master thesis [Galvão 2012] of the author. Chapter 2 serves as a prelude to the other chapters. For instance, Section 2.2 introduces the basic concepts used throughout the thesis and provides a literature review of several variants of the permutation sorting problem.

Chapter 3 corresponds to an article [Dias et al. 2014] published in *Journal of Bioinformatics and Computational Biology*. In this chapter, we present a general heuristic for permutation sorting problems. The heuristic works by iteratively improving an initial solution produced by other algorithm. In each step, it makes a local change within a sliding window, which moves across the solution. The main idea employed by the heuristic is to transform the sliding window into a small instance of the permutation sorting problem

in such a way that an optimal solution for that instance can be retrieved from a solution database. To evaluate the heuristic, we applied it to the solutions provided by 23 approximation algorithms. The performance of the heuristic varied considerably depending on the algorithm that produced the initial solutions: it ranged from almost 5% of improvement to near 100%. The observed variation is mainly due to the quality of the initial solutions: the closer they are to the optimal solution, the more difficult it is to improve them.

Chapter 4 corresponds to an article [Galvão and Dias 2014] published in *Journal of Universal Computer Science*. The best known algorithms for the problem of sorting by transpositions are based on a standard tool for tackling permutation sorting problems, the cycle graph. In an attempt to bypass it, a few researches proposed algorithms based on alternative tools. In Chapter 4, we address three of these algorithms: a 2.25-approximation algorithm proposed by Walter et al. (2000), a 3-approximation algorithm proposed by Benoît-Gagné and Hamel (2007), and a heuristic proposed by Guyer et al. (1997). On the theoretical side, we close a missing gap on the proof of the approximation ratio of Benoît-Gagné and Hamel's algorithm [Benoît-Gagné and Hamel 2007] and we demonstrate a way to run their algorithm in $O(n \log n)$ time. Moreover, we propose a minor adaptation to Guyer, Heath, and Vergara's heuristic [Guyer et al. 1997] that allow us to prove an approximation bound of 3. On the evaluation side, we present experimental data indicating that Walter, Dias, and Meidanis' algorithm [Walter et al. 2000] is the best of the algorithms based on alternative approaches and that it is the only one comparable to the algorithms based on the cycle graph.

Chapter 5 corresponds to an article [Galvão et al. 2015b] published in *Algorithms for Molecular Biology*. In this chapter, we investigate the problem of sorting a signed permutation by short operations (*i. e.* operations involving just few genes). The biological relevance of this problem is grounded on the assumption that rearrangement events affecting large portions of a genome are less likely to occur. In the past, corroborating evidence has emerged, that is, separate sets of observations have shown the prevalence and significance of short reversals in the evolution of bacterial genomes [Dalevi et al. 2002, Lefebvre et al. 2003] and lower eukaryote genomes [McLysaght et al. 2000, Seoighe et al. 2000]. This fact, together with the realization that signed permutations constitute a more biologically relevant model for genomes, motivated us to investigate the problem of sorting a signed permutation by short operations.

In preliminary work, we [Galvão and Dias 2014a] investigated the problem of sorting a signed permutation by reversals of length at most 3 and presented three approximation algorithms, the best one having an approximation factor of 9. In Chapter 5, we not only present an approximation algorithm with a better approximation factor, but also consider other variants of the problem. More precisely, we study four variants of the permutation sorting problem: (i) the problem of sorting a signed permutation by reversals of length at most 2, (ii) the problem of sorting a signed permutation by reversals of length at most 3, (iii) the problem of sorting a signed permutation by reversals and transpositions of length at most 2, and (iv) the problem of sorting a signed permutation by reversals and transpositions of length at most 3. We present polynomial-time solutions for problems (i) and (iii), a 5-approximation for problem (ii), and a 3-approximation for problem (iv). Moreover, we show that the expected approximation factor of the 5-approximation algo-

rithm is not greater than 3 for random signed permutations with more than 12 elements. Finally, we present experimental results that show that the approximation factors of the approximation algorithms cannot be smaller than 3. In particular, this means that the approximation factor of the 3-approximation algorithm is tight.

Chapter 6 corresponds to an article [Galvão et al. 2015a] published in the proceedings of the *11th International Symposium on Bioinformatics Research and Applications*. In this chapter, we consider the problem of sorting a circular permutation by reversals of length at most 2. Polynomial-time solutions for the unsigned version of this problem were known, but the signed version remained open. In Chapter 6, we present the first polynomial-time solution for the signed version of the problem. Moreover, we perform an experiment to infer distances and phylogenies for published *Yersinia* genomes and compare the results with the phylogenies presented in previous works [Darling et al. 2008, Egri-Nagy et al. 2014].

Recently, the full version of the conference paper corresponding to Chapter 6 has been accepted for publication in *IEEE/ACM Transactions on Computational Biology and Bioinformatics* [Galvão et al. 2016]. The first relevant difference between the journal version and the conference version of the paper is that the journal version contains expanded explanations and expositions of the methods. In fact, the theoretical sections have been rewritten, at least to some extent, to provide more details. In particular, the journal version uses some of the formalism introduced by Meidanis et al. (2000) and Solomon et al. (2003) for dealing with the problem of sorting circular permutations by reversals. The second relevant difference is that the journal version shows additional experimental results: in addition to the experiment with *Yersinia* genomes, we also performed an experiment to infer distances and phylogenies of γ-proteobacterial genomes. Lastly, the third relevant difference is that the journal version presents a web tool for rearrangement-based phylogenetic inference using short operations. The aim of this tool is to facilitate phylogenetic studies based on the methods proposed in the paper. It is available at:

`http://mirza.ic.unicamp.br:8080/shortphy`.

3. Conclusion

In this paper, we presented a summary of the PhD thesis defended by the author [Galvão 2015]. The thesis is focused on solving variants of the permutation sorting problem that involves reversals or transpositions. The main contributions of the thesis include: i) a general heuristic that can be used to improve the solutions provided by any non-optimal algorithm for the permutation sorting problem; ii) a theoretical and experimental investigation of three algorithms based on alternative approaches for the problem of sorting by transpositions; iii) exact and approximation algorithms for the problem of sorting signed permutations by short operations; and iv) a web tool for rearrangement-based phylogenetic inference using short operations. All contributions were published in peer-reviewed international journals, such as described in Table 1.

Table 1. Overview of the bibliographic production. The third column shows the 2014 JCR Impact Factor of the journals, while the fourth column shows the 2014 Qualis classification in Computer Science. The journal *Algorithms for Molecular Biology* does not have a 2014 (nor a 2013) Qualis classification in Computer Science, hence we are showing the 2012 Qualis classification.

Chap.	Journal	JCR	Qualis
2	ACM Journal of Experimental Algorithmics	–	B4
3	Journal of Bioinformatics and Computational Biology	0.78	B1
4	Journal of Universal Computer Science	0.46	B1
5	Algorithms for Molecular Biology	1.46	A1
6	IEEE/ACM Transactions on Computational Biology and Bioinformatics	1.43	B1

References

Benoît-Gagné, M. and Hamel, S. (2007). A new and faster method of sorting by transpositions. In *Proceedings of the 18th Annual Symposium on Combinatorial Pattern Matching (CPM'2007)*, volume 4580 of *Lecture Notes in Computer Science*, pages 131–141, London, Ontario, Canada. Springer-Verlag.

Dalevi, D. A., Eriksen, N., Eriksson, K., and Andersson, S. G. E. (2002). Measuring genome divergence in bacteria: A case study using chlamydian data. *Journal of Molecular Evolution*, 55(1):24–36.

Darling, A. E., Miklós, I., and Ragan, M. A. (2008). Dynamics of genome rearrangement in bacterial populations. *PLoS Genetics*, 4(7):e1000128.

Dias, U., Galvão, G. R., Lintzmayer, C. N., and Dias, Z. (2014). A general heuristic for genome rearrangement problems. *Journal of Bioinformatics and Computational Biology*, 12(3):1450012.

Egri-Nagy, A., Gebhardt, V., Tanaka, M. M., and Francis, A. R. (2014). Group-theoretic models of the inversion process in bacterial genomes. *Journal of Mathematical Biology*, 69(1):243–265.

Fertin, G., Labarre, A., Rusu, I., Tannier, E., and Vialette, S. (2009). *Combinatorics of Genome Rearrangements*. The MIT Press, Cambridge, MA, USA.

Galvão, G. R. (2012). *Uma Ferramenta de Auditoria para Algoritmos de Rearranjo de Genomas*. Master's thesis, University of Campinas. In Portuguese.

Galvão, G. R. (2015). *Algorithms for Sorting by Reversals or Transpositions, with Application to Genome Rearrangement*. PhD thesis, University of Campinas.

Galvão, G. R., Baudet, C., and Dias, Z. (2015a). Sorting signed circular permutations by super short reversals. In *Proceedings of the 11th International Symposium on Bioinformatics Research and Applications (ISBRA'2015)*, volume 9096 of *Lecture Notes in Computer Science*, pages 272–283. Springer International Publishing.

Galvão, G. R., Baudet, C., and Dias, Z. (2016). Sorting circular permutations by super short reversals. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*. To appear.

Galvão, G. R. and Dias, Z. (2014a). Approximation algorithms for sorting by signed short reversals. In *Proceedings of the 5th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics (ACM-BCB'2014)*, pages 360–369, Newport Beach, California, USA. ACM Press.

Galvão, G. R. and Dias, Z. (2014b). An audit tool for genome rearrangement algorithms. *ACM Journal of Experimental Algorithmics*, 19:1.1–1.34.

Galvão, G. R., Lee, O., and Dias, Z. (2015b). Sorting signed permutations by short operations. *Algorithms for Molecular Biology*, 10(12).

Galvão, G. R. and Dias, Z. (2014). On alternative approaches for approximating the transposition distance. *Journal of Universal Computer Science*, 20(9):1259–1283.

Gascuel, O. (2005). *Mathematics of Evolution and Phylogeny*. Oxford University Press, Inc., New York, NY, USA.

Guyer, S. A., Heath, L. S., and Vergara, J. P. C. (1997). Subsequence and run heuristics for sorting by transpositions. Technical Report TR-97-20, Virginia Polytechnic Institute & State University.

Lefebvre, J. F., El-Mabrouk, N., Tillier, E., and Sankoff, D. (2003). Detection and validation of single gene inversions. *Bioinformatics*, 19(suppl 1):i190–i196.

Lemey, P., Salemi, M., and Vandamme, A. (2009). *The Phylogenetic Handbook: A Practical Approach to Phylogenetic Analysis and Hypothesis Testing*. Cambridge University Press, Cambridge, UK.

McLysaght, A., Seoighe, C., and Wolfe, K. H. (2000). High frequency of inversions during eukaryote gene order evolution. In Sankoff, D. and Nadeau, J. H., editors, *Comparative Genomics*, volume 1 of *Computational Biology*, pages 47–58. Springer Netherlands.

Meidanis, J., Walter, M. E. M. T., and Dias, Z. (2000). Reversal distance of signed circular chromosomes. Technical Report IC-00-23, Institute of Computing, University of Campinas.

Saitou, N. and Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, 4(1):406–425.

Seoighe, C., Federspiel, N., Jones, T., Hansen, N., Bivolarovic, V., Surzycki, R., Tamse, R., Komp, C., Huizar, L., Davis, R. W., Scherer, S., Tait, E., Shaw, D. J., Harris, D., Murphy, L., Oliver, K., Taylor, K., Rajandream, M. A., Barrell, B. G., and Wolfe, K. H. (2000). Prevalence of small inversions in yeast gene order evolution. *Proceedings of the National Academy of Sciences USA*, 97(26):14433–14437.

Solomon, A., Sutcliffe, P., and Lister, R. (2003). Sorting circular permutations by reversal. In Dehne, F., Sack, J.-R., and Smid, M., editors, *Algorithms and Data Structures*, volume 2748 of *LNCS*, pages 319–328. Springer Berlin Heidelberg.

Walter, M. E. M. T., Dias, Z., and Meidanis, J. (2000). A new approach for approximating the transposition distance. In *Proceedings of the Seventh International Symposium on String Processing Information Retrieval (SPIRE'2000)*, pages 199–208, Washington, DC, USA. IEEE Computer Society.