RESEARCH PROJECT

ALGORITHMIC QUESTIONS ON MOLECULAR BIOLOGY

C.E. FERREIRA AND M.-F. SAGOT

ABSTRACT. The research proposed here will mainly focus on computational biology. The technological advances in biology related areas allowed a huge amount of information to be extracted. It is one of the main current challenges for biologists to be able to extract knowledge from all this data and a challenge also for the computer scientists to help the biologists in this process. The analysis of this data involves, among many other issues, the formalization of problems and the search for efficient algorithms to solve these problems. This is a continuous process in the sense that the solutions produced by the algorithms will usually lead a refinement of the models and problem formulations, and will therefore require new algorithms. The main goal of this project is to work on these types of questions: the formalization of problems coming from biology and the search for efficient algorithms for these problems.

1. INTRODUCTION AND JUSTIFICATION

Living organisms are extraordinarily complex systems which recent technological advances have enabled to start studying in a completely new fashion, at scales never contemplated before. Although mathematical models have since a long time been attempted for many biological phenomena, usually this was at a relatively small specialized level. Even then, the data was often missing to test, revise and refine such models. This data is now coming so fast that it is the models and their accuracy that are lagging increasingly behind. The coverage of the data is also extensive enough already that one may start considering to model larger and wider covering systems. This is totally essential if one hopes to get a truly good understanding of how complex organisms function and reproduce.

Although the speed of arrival of new data, and their overall size have often been mentioned as the main bottlenecks to the mathematical and computational modeling and treatment of such new information, the real bottleneck is situated elsewhere. It is in the relative lack of sophistication of currently existing computational tools for dealing even sometimes with small biological datasets. The mathematical and algorithmic theory is often not yet there that would enable to fully satisfactorily treat the complexity of living organisms.

This project addresses exactly this point. The French and the Brazilian groups gather people with interest in mainly three areas of computer science: computational biology, combinatorial optimization and graph theory. These areas are related to each other, as many problems in the first two are modeled and formalized as problems in graph theory. Also, many problems in computational biology are in fact optimization problems. Sharing knowledge within these areas is therefore an important step towards solving important practical problems, as well as towards developing the background to find more efficient solutions for such problems. We want to use the theoretical skills of both groups to improve some of the mathematical and algorithmical theory needed to get a better understanding of evolution and of the preservation of structures in living systems. We intend to work on four main topics: sequence analysis and models for molecular evolution (Part 1 of this project), large-scale comparative genomics and genome dynamics (Part 2), phylogeny (Part 3), and biochemical networks (Part 4). Each of these topics is already under investigation by one or both sides, either independently or in collaboration.

Part 1: Sequence analysis and models for molecular evolution

This part is basically concerned with identifying motifs of different types and sizes in molecular sequences. This identification requires also to have good models of the evolution of such sequences and this is therefore a third topic that appears as background to the other two.

The following subtopics will more particularly interest us in the context of this project.

1.1: Improving or investigating the various models for motifs that have been developed over the years. The motifs here include: complex motifs, that is motifs which may be composed of various parts separated by more or less constrained distances; footprint motifs, that is motifs that explicitly take a (known) phylogeny into account when trying to infer motifs from a set of orthologous sequences; finally, motifs that try to combine both the pattern and position weight matrix models for representing motifs.

1.2: Systematically investigating new approaches for motif inference, in particular by exploring probabilistic methods and general machine learning techniques, and more specifically analyzing the feasibility of non supervised motifs learning using biclustering techniques adapted to motifs.

1.3: Extending the types of motifs that have been considered to treat long motifs (which may require filtering techniques specially adapted to the problem – genes may be seen as very long motifs, split in the case of eukaryotes), motifs under an edit distance model, parameterized motifs (motifs which match under general functions), motifs appearing in permuted forms.

1.4: Systematically investigating new indexes for old and new motif models, in particular indexes for approximate motifs search and inference and for complex motifs.

1.5: Analyzing various exact or approximate formulations of motif clustering, from the most elegant ones (such as bases of motifs already undergoing investigation in one of the groups) to heuristical clustering methods.

1.6: Improving the algorithms for multiple sequence alignment that underly motif search or inference methods, in particular, exploring new cost functions such as normalized edit distance (where the alignment quality of the sequences is measured by considering the average of the quality of the local alignment of the sequences) [2, 3, 24, 28, 35].

1.7: Exploiting conservation differences, that is, differences in evolutionary constraints, to identify gene regulatory sequences. Algorithms will be considered for scanning sequences from different organisms for conserved equivalent segments and within these segments, to identify hyper conserved and hyper variable regions.

Part 2: Large-scale comparative genomics and genome dynamics

It has since long been known that genomes are not static. The work of Barbara Clintock in the late 40's showing that genes could jump spontaneously from one site to another was a first clear sign of this. "Jumping genes" were called transposable elements by Clintock. Genes may also get duplicated. There are strong indications that the duplication may sometimes affect whole chromosomes or even genomes or, inversely, only pieces of a gene, in particular exons. It has thus been shown that, during evolution, DNA segments coding for modules or domains in proteins have been duplicated and rearranged through what has been called intronic recombination. By shuffling modules between genes, protein families have thus evolved. Genomic segments can be reversed, in general through ectopic recombination, or deleted. Chromosomes in multi-chromosomal organisms may undergo fusion or fission, or exchange genetic material with another chromosome (through homologous recombination), usually at their ends (translocation) or internally. Genetic material may also be transferred across sub-species or species (lateral transfer), thus leading to the insertion of new elements in a genome. Parts of a genome may be amplified, through, for instance, slippage resulting in the multiplication of the copies of a tandem repeat.

Although much is known about the dynamic behavior of genomes, much more remains to be discovered about the forces and exact mechanisms behind such dynamics, its function and extend, the frequency of each type of rearrangement, and the impact genomic reorganizations may have on gene expression and genome development.

Below follows a list of topics where, we believe, the collaboration within this project might yield significant breakthroughs.

2.1: Algorithms and complexity analysis for calculating a rearrangement distance between two or more genomes under various models. Classical methods of DNA sequence comparison assumed that sequences may only mutate by operations that act on individual nucleotides, *i.e.*, substitutions, insertions, and deletions. More recently, additional studies considered large scale genome rearrangement events such as inversions [11, 12, 17, 18, 33], transpositions [13, 14, 19] and translocation. We aim to broaden the theory of genome rearrangement in several directions, and to tighten the contact between the theoretical analysis and the real data that are gradually becoming available. The key topics we shall study are algorithms for sorting by signed reversals, length-sensitive sorting by reversal [5], sorting by transpositions, handling duplicated genes, handling missing genes, handling multiple genomes.

2.2: Modeling, detection and analysis of "segments conserved by rearrangements" [6, 15]. "Segments conserved by rearrangements" mean parts of a chromosome which are relatively stable under large-scale evolutionary events. Looking for such segments is a difficult task for two reasons. The first one is that the conservation is not exact. Some rearrangements preserve the function associated with a segment provided there are "not many" of them. How precisely to define the number and type that should be allowed, and therefore which definition(s) to adopt for a conserved segment (possibly there will be more than one depending on the biological question) remains very much an open problem. Our first task will thus be to derive models that are satisfying both mathematically and biologically. The second difficulty of the problem is that such models may be hard to compute.

2.3: Study of breakpoint regions of the genome [4, 7, 34]. This consists in analyzing the regions where rearrangements have broken the genome, and trying to find some characteristics that may enable to classify them according to the type of rearrangement that gave them origin. The characteristics sought could be the motifs or repeats such regions may contain, or some other features still to be determined. We intend to build methods for detecting such regions as accurately as possible (this is the counterpart of the conserved segments mentioned just above). Then by studying the distribution and length of these regions, we shall try to evaluate the reality of the "fragile regions" model, which asserts that there are evolution hotspots in the genome. This theory is under discussion in the scientific community, and still lacks efficient theoretical bases. We shall combine gene homology data and global genomic alignment data and provide reliable tools and analyses based on previous studies on rearrangements.

2.4: We have interest also in a particular subproblem of the previous one, namely the problem of alignments with inversions [11, 12, 17, 18, 33]. Sequence alignments are broadly studied for biological sequence comparison but considering only biological events such as mutations, insertions and deletions. Other biological events such as inversions are not automatically detected by the usual alignment algorithms. Some alternative strategies have been considered in the attempt to include inversions and other types of rearrangements. We plan to improve further on some substantial results we have already concerning this topic [9], and possibly to generalize them to other types of alignments and objective functions.

2.5: Repetitions, recombinations and rearrangements. The objective is to design algorithms for identifying various types of repeats, comparing various alleles of a tandem repeat and studying the relationship between repeats and recombination. We shall investigate new models, algorithms and indexes for identifying various types of repeats in a sequence. The work will start by attempting a typology of the various types of repeats that may be found in biological sequences. Mathematical models and efficient algorithms for their detection will then be investigated for some of these repeats. A tandem repeat has a history and any two individuals may have different tandem repeat sequences at the same genomic location. For various scientific reasons, biologists are interested in tracing back the history of a tandem repeat and in comparing different alleles of a tandem repeat. In the proposed work, we shall focus on combinatorial and algorithmic aspects of the duplication phenomenon. The more complicated case where recombination is present between the copies of a tandem repeat will also be addressed.

Part 3: Phylogeny

Although rearrangements are now generally known to be a major force of the evolution of organisms, they have been surprisingly little used to study such evolution, in particular to infer phylogenies. Phylogenies have instead continued to be most often derived from the point mutation information of one or more genes. Although it is well known, and is indeed a current topic of investigation, that the "evolutionary story" told by a gene often contradicts the "story" told by another gene, or the "story" obtained by using some more general information from an organism (*e.g.* a set of genes taken together, or a set of morphological characters), such evolutionary stories continue to be represented in a hierarchical manner, that is using trees as models instead of graphs [8, 20, 21, 25].

This difficult problem will be addressed in a step-wise manner by considering the following subproblems.

3.1: Revisiting the problem of a phylogenetic tree reconstruction. When building phylogenetic trees from molecular sequences, a natural model that captures the problem is a variant of the Steiner tree problem, where we allow terminals to be only leaves of the tree [10, 16, 26, 27]. Another variant of this problem that seems particularly relevant for our purposes is one where we are given also a permutation of the terminals and we want a minimum-cost Steiner tree with terminal leaves that respects the given permutation. By this we mean that, if in the permutation terminals r_1, r_2, r_3 and r_4 appear in this order, then the paths in the tree between r_1 and r_3 and between r_2 and r_4 intersect. We have some preliminary results on these problems and intend to push them further. The second problem within this context concerns the maximum homeomorphic agreement subtree problem, which has several applications in the computation of a distance/dissimilarity between two trees. It consists of the following: given a collection of rooted trees with its leaves labelled by elements of a set A, find a subset of A of maximum cardinality such that the subtrees of the given trees "induced" by the leaves with labels in this subset coincide, except for possible edge subdivisions.

3.2: Computing a recombination distance between two phylogenetic trees [20, 21, 22, 23]. In our preliminary theoretical work with this topic, the distance we explored is based on the MAF ("Maximum Agreement Forest") between two trees [31]. We shall continue this work by examining the relation between the MAF and another possibly more appropriate recombination distance between trees: the SPR ("Subtree Prune and Regraft") distance. This is the minimum number of subtree cuts and regrafts one must do in order to transform one tree into another. We shall also work on elaborating new algorithms for the MAF and other distances, either with a better ratio than those established so far, including by us, or parameterized.

3.3: Exploring the network (graph) nature of the evolution of organisms by combining optimization methods and graph algorithms for inferring what are called reticulate phylogenies.

Part 4: Biochemical networks

It is now commonly accepted that the functioning and development of a living organism is controlled by the networks of interactions between its genes, proteins, and small molecules. Studying such networks and their underlying complexity is the main objective of this part. This objective hides a second one, no less crucial, which is to greatly improve the mathematical and algorithmical theory needed to accurately model, and then explore and analyze highly complex living systems. Biochemical networks may represent protein-protein interactions, the metabolism of an organism, its system of gene expression regulation, or even, mixed networks that contain information coming from various of the previous sources.

The amount and spread of the data now becoming available enable us also to introduce an evolutionary perspective into the study of living organisms, and in particular of biochemical networks. Evolution is a general underlying principle of life that allows us to compare and decipher the meaning and function of structure, the modification of biochemical pathways and networks, the preservation and variation of cell signalling systems, and so on. It thus serves to study the fundamental aspects of life, taking advantage for this of the exploratory and comparative possibilities provided, in particular, by the availability of an increasing number of whole sequences and datasets from different genomes.

The main topic for which the intensive collaboration among the members of this project can be profitable is the search of pathways, motifs and modules in different kinds of biochemical networks, that is, the search of what is functional, and therefore preserved by evolution. In particular, we shall be concerned with the following two main topics.

4.1: Motifs and modules in biochemical networks [1, 29]. No fully satisfying or complete definition of motifs and modules in biochemical networks exist and most of the work will consist in exploring the various which may be considered (topological or other) and the algorithmical complexity of such definitions. For each, efficient data structures, filters and algorithms for both searching known motifs and for inferring new ones in large networks will be developed. The definitions will of course vary depending on the type of biochemical network that is considered.

The question of the statistical significance of the motifs identified will be of primary importance. This question is still open. An answer to it may depend on the definition of a random graph that is appropriate to the biological problem at hand, a definition of a motif occurrence in such a network, and how to calculate the probability of such motifs. The possibility to transpose questions and results on motif statistics in a random sequence to motifs statistics in a random network will be examined. This will be a more exploratory research activity.

Finally, modules and motifs will also be essential instruments for studying and understanding the evolution of networks. Indeed metabolic pathways have already been used to infer phylogenies but the topological aspects of such pathways are only very partially and indirectly taken into account. This is another area that will be developed, possibly in conjunction with other types of information.

4.2: Reconstruction, alignment and simulation of metabolic pathways. Metabolic pathways reflect the sum of an organism's chemical reactions, and their elucidation is key to the understanding of cellular processes as a whole. Such pathways can be represented as labeled graphs and networks of processes, thus making them amenable to algorithmic analyses of several kinds. Our objective is to combine methods for computational analysis and simulation of these structures with experimental work that reveals the (kinetic and other) parameters that are required to characterize the behavior of these systems in order to allow life science researchers to better understand how metabolic pathways function.

In our work, we aim to provide researchers with systematic and predictive means to do their work. These include the ability to compare metabolisms both of a variety of organisms as well as of similar processes within the same organism, the provision of tools and methods to do both static and dynamic analyses of pathways, and the ability to reconstruct complex pathways from their constituents. Note that some of the methods developed in this context are applicable also to other cases, such as regulatory networks, or protein-protein interaction networks.

1.1. **Justification.** The two research groups involved in this project are complementary in the following sense. The French group is strongly active in the area of computational biology. Most of its members come from computer science or mathematics, but some have majors in biology or bioinformatics. They are also actively in contact with biologists and maintain a strong collaboration with researchers in molecular biology. The coordinator of the French side, Dr. Marie-France Sagot, has a background in computer science, more specifically algorithmics. She worked for four years at the Pasteur Institute in Paris and is now a Director of Research at the INRIA. She is physically located at the Claude Bernard University in Lyon, and belongs to the Laboratoire of Biometry and Evolutionary Biology of the University. The Brazilian group is strongly active in combinatorial optimization, computational complexity and development of algorithms in general, and has, in the past years, invested in the area of computational biology. All but one of them are members of the Combinatorics and Combinatorial Optimization Research Group, at the Department of Computer Science of the University of São Paulo, Brazil.

This collaboration will be very beneficial for both sides. The Brazilian members will benefit from the expertise in computational biology from the French members while the latter will be able to count on the general experience in combinatorial optimization, computational complexity and algorithmics of the Brazilian members to reinforce these aspects inside the French group. The two groups count with a number of PhD students that will also benefit from this collaboration.

It is important to note that the collaboration has already existed for many years now. Dr. Marie-France Sagot was a student at the University of São Paulo before coming to France for the DÉA. She kept her contacts at USP. Such contacts lead to an official collaboration through a CAPES-COFECUB project from 1999 to 2001, coordinated by Dr. Sagot and Prof. Yoshiko Wakabayashi (USP), also a participant of the current proposal. Prof. Wakabayashi and Dr. Sagot have been collaborating for many years. Recently, Springer-Verlag published a book named *Recent Advances in Algorithms and Combinatorics*, edited by B. Reed and C.L. Sales (CMS Books in Mathematics, 2003), with several chapters. One of them, *Pattern inference under many guises* [32], was written by the two of them, and addresses part of the topics that will be under investigation within this project. One of the members of the Brazilian group, when a PhD student, spent 9 months in a research mission under the supervision of the French coordinator (two papers, one accepted and one submitted, resulted from the mission [31, 30]), and recently another Brazilian student, who is now finishing his PhD, did the same. Currently a member of the Brazilian group is in a research visit with the French partners for a period of 6 weeks. This visit is being funded (informally) by the INRIA (stay) and the ProNEx 107/97 - MCT/CNPq project (Proc. CNPq 664107/1997-4, plane tickets).

We believe that, with this new project, we shall intensify the already existing collaboration and give continuity to the student interchanges, which are so important for their training as it exposes them to an international cooperation.

2. Main goals

Recently, this same group has submitted a proposal to INRIA of an Associated Team between the Combinatorics and Combinatorial Optimization Research Group (USP, Brazil) and the HELIX group (Lyon/Grenoble, France).

Our main goal is to strengthen and start new collaborations among the participants of the project and make progress on the resolution of the problems listed above.

Another goal is to give to our PhD students an opportunity to take part of an international cooperation, to initiate international contacts, to be exposed to a different research environment, and to benefit from the expertise and the experience of different researchers.

3. Work plan and schedule preview

For the first year, we shall concentrate our efforts in the following points from the list presented above.

1.3 Inferring long motifs, in particular for gene prediction.

1.5 Analyzing exact formulations of motif clustering, more particularly bases of motifs.

1.6 Improving the algorithms for multiple alignment, in general and for the purpose of predicting eukaryotic genes.

1.7 Exploring conservation differences to identify gene regulatory sequences, in general and for the purpose of predicting eukaryotic genes.

2.1 Studying algorithms and analyzing the complexity for calculating a rearrangement distance between two or more genomes under various models.

2.2 Modeling, detecting and analyzing “segments conserved by rearrangements”.

2.4 Continuing the work on alignments with inversions.

3.1 Revisiting the problem of a phylogenetic tree reconstruction by examining two variants (Steiner trees with terminal leaves and maximum homeomorphic agreement subtree problems).

3.2 Continuing the work on computing a recombination distance between two phylogenetic trees.

4.1 Searching for and inferring motifs and modules in biochemical networks.

To stimulate the collaboration between the two groups, we plan visits of members of the project and some of their PhD students to the partner institution. These scientific missions will provide the conditions for the interaction between the two groups to occur.

Specifically, as activities within this project and within the Associated Teams proposal, we are planning to organize one or two meeting a year, involving members of both groups.

Meetings of this kind have been organized by the Combinatorics and Combinatorial Optimization Research Group for the past years as part of the ProNEx project – Complexity of Discrete Structures (`http://www.ime.usp.br/~yoshi/pronex/`), and have shown to be quite effective.

For 2005, we are planning a smaller meeting in April and one a bit larger in November, both to be held in Brazil. Similar meetings shall happen in the second year of the project. The French missions are going to concentrate around these meetings.

Besides these meetings, we are planning some missions from the Brazilian researcher to France. Specifically, the following Brazilian participants of the project plan a mission for 2005:

1. Alair Pereira do Lago, July or December, for 5 weeks;
2. Cristina G. Fernandes, July 2005, for two weeks;
3. Estela Maris Rodrigues, July or December 2005, for 4 weeks.

For 2006, we predict two missions from researchers, and one from a PhD student, from Brazil to France.

The schedule of the missions for the French side consists of the following for 2005.

1. Marie-France Sagot, April, for 1 week, and November, for 1 or 2 weeks;
2. Eric Tannier, April, for 1 week, and November, for 1 or 2 weeks;
3. Laurent Guéguen, November, for 1 or 2 weeks;
4. Christian Gautier, November, for 1 or 2 weeks.

Possibly some of these stays might be extended for a longer period, if that seems feasible and interesting for the project.

The plans for the second year from the French side are similar, consisting of around 4 visits from researchers from France to Brazil.

4. Available infrastructure and methodology

For the past three decades, the Department of Computer Science at IME-USP has made a great effort in the areas of Combinatorial Optimization, Combinatorics and Theoretical Computer Science. In terms of research, it is one of the main centers in these areas in Brazil.

As for the French group, they are located at the Claude Bernard University in Lyon, and belong to the Laboratoire of Biometry and Evolutionary Biology of the University.

Both institutions have good infrastructure for the development of the project: computer laboratories, libraries and offices.

As for methodology, as mentioned in the previous section, we plan visits from members of one of the groups to the other, and one or two meetings, where the participants shall work together on some of the problems mentioned above.

We also intend to take part in good conferences of the area, which have a very strict policy for the selection of the presented works. Such participations, besides helping disseminate the work that is being done within the project, may bring some feedback on the work from the academic community.

Each of the groups is responsible for a weekly seminar where on-going research is presented and discussed. The seminars are attended by the members of this project, besides other faculty members and students from related areas. Naturally the results obtained in this project will be presented at these seminars.

As with other projects, we would maintain a webpage with information on the project (participant institutions, members, project proposal, reports, publications, meetings, etc).

5. Methods of analyzing the results

As concrete results, we expect publications in scientific journals, proceedings of international well-known conferences and participations in some of the main workshops in the area. We also expect that some students will get their degree working on topics within the scope of this project.

References

1. E. Alm and A. Arkin, *Biological networks*, Current opinion in structural biology **13** (2003), 193–202.
2. A.N. Arslan and O. Egecioglu, *An efficient uniform-cost normalized edit distance algorithm*, Tech. Report TRCS99-14, Departament of Computer Science - University of California, Santa Barbara, 24, 1999.
3. ______, *Efficient algorithms for normalized edit distance*, J. Discret. Algorithms **1** (2000), no. 1, 3–20.
4. J.A. Bailey, R. Baertsch, W.J. Kent, D. Haussler, and E.E. Eichler, *Hotspots of mammalian chromosomal evolution*, Genome Biol. **5** (2004), no. 4, 7pp.
5. M.A. Bender, D. Ge, S. He, H. Hu, R.Y. Pinter, S. Skiena, and F. Swidan, *Improved bounds on sorting with length-weighted reversals*, Proceedings of the fifteenth annual ACM-SIAM symposium on Discrete algorithms, Society for Industrial and Applied Mathematics, 2004, pp. 919–928.
6. S. Bérard, A. Bergeron, and C. Chauve, *Conserved structures in evolution*, 2nd RECOMB Comparative Genomics Satellite Workshop (2004), 14pp, to appear.
7. G. Bourque, P.A. Pevzner, and G. Tesler, *Reconstructing the genomic architecture of ancestral mammals: Lessons from human, mouse, and rat genomes*, Genome Research **14** (2004), no. 4, 507–516.
8. K.A. Crandall, *Intraspecific phylogenetics: support for dental transmission of human immunodeficiency virus*, J. of Virology **69** (1995), no. 1, 2351–2356.
9. A.P. do Lago, I. Muchnika, and C.A. Kulikowski, *A sparse dynamic programming algorithm for alignment with non-overlapping inversions*, submitted, 2003.
10. D. E. Drake and S. Hougardy, *On approximation algorithms for the terminal Steiner tree problem*, Information Processing Letters **89** (2004), no. 1, 15–18.
11. N. El-Mabrouk, *Genome rearrangement by reversals and insertions/deletions of contiguous segments*, Combinatorial pattern matching (Montreal, QC, 2000), Lecture Notes in Comput. Sci., vol. 1848, Springer, Berlin, 2000, pp. 222–234.

12. ———, *Sorting signed permutations by reversals and insertions/deletions of contiguous segments*, J. Discrete Algorithms **1** (2000), no. 1, 105–121.
13. N. Eriksen, *(1 + ϵ)-approximation of sorting by reversals and transpositions*, Theor. Comput. Sci. **289** (2002), no. 1, 517–529.
14. H. Eriksson, K. Eriksson, J. Karlander, L. Svensson, and J. Wastlund, *Sorting a bridge hand*, Discrete Math. **241** (2001), no. 1–3, 289–300.
15. M. Figeac and J.-S. Varré, *Sorting by reversals with common intervals*, Algorithms in Bioinformatics: 4th International Workshop, WABI 2004, LNCS **3240** (2004), 26–35.
16. B. Fuchs, *A note on the terminal Steiner tree problem*, Information Processing Letters **87** (2003), 219–220.
17. S. Hannenhalli and P.A. Pevzner, *Transforming cabbage into turnip: polynomial algorithm for sorting signed permutations by reversals*, ACM Symposium on Theory of Computing, Association for Computing Machinery, 1995, pp. 178–189.
18. ———, *Transforming cabbage into turnip: polynomial algorithm for sorting signed permutations by reversals*, J. ACM **46** (1999), no. 1, 1–27.
19. T. Hartman and R. Shamir, *A simpler* 1.5*-approximation algorithm for sorting by transpositions*, Proceedings of CPM'03, LNCS **2676** (2003), 156–169.
20. J. Hein, *Reconstructing the evolution of sequences subject to recombination using parsimony*, Mathematical Biosciences **98** (1990), 185–200.
21. ———, *A heuristic method to reconstruct the history of sequences subject to recombination*, J. of Mol. Evol. **36** (1993), 396–405.
22. J. Hein, T. Jiang, L. Wang, and K. Zhang, *On the complexity of comparing evolutionary trees*, Proceedings of the 6th Annual Symposium on Combinatorial Pattern Matching (Springer-Verlag, ed.), Lecture Notes in Computer Science, vol. 937, 1995.
23. ———, *On the complexity of comparing evolutionary trees*, Discrete Appl. Math. **71** (1996), no. 1-3, 153–169.
24. B. John Oommen and K. Zhang, *The normalized string editing problem revisited*, IEEE Transactions on Pattern Analysis and Machine Intelligence **18** (1996), no. 6, 669–672.
25. S.K. Kannan and T. Warnow, *Inferring evolutionary history from dna sequences*, SIAM J. on Computing **23** (1994), 713–737.
26. G. Lin and G. Xue, *On the terminal Steiner tree problem*, Information Processing Letters **84** (2002), 103–107.
27. C. L. Lu, C. Y. Tang, and R. C.-T. Lee, *The full Steiner tree problem*, Theoretical Computer Science **306** (2003), 55–67.
28. A. Marzal and E. Vidal, *Computation of normalized edit distance and applications*, IEEE Transactions on Pattern Analysis and Machine Intelligence **15** (1993), no. 9, 926–932.

29. R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon, *Network motifs: Simple building blocks of complex networks*, Science **298** (2002), 824–827.

30. E.M. Rodrigues, M.-F. Sagot, and Y. Wakabayashi, *The maximum agreement forest problem: approximation algorithms and computational experiments*, submitted.

31. ______, *Some approximation results for the maximum agreement forest problem*, Proceedings of APPROX-RANDOM 2001 (Approximation, Randomization, and Combinatorial Optimization: Algorithms and Techniques) (M. Goemans and K. Jansen abd J.D.P. Rolim abd L. Trevisan, eds.), Lecture Notes in Computer Siences, vol. 2129, Springer-Verlag, 2001, pp. 159–169.

32. M.-F. Sagot and Y. Wakabayashi, *Pattern inference under many guises*, Recent advances in algorithms and combinatorics, CMS Books Math./Ouvrages Math. SMC, vol. 11, Springer, New York, 2003, pp. 245–287.

33. M. Schöniger and M. S. Waterman, *A local algorithm for DNA sequence alignment with inversions.*, Bulletin of Mathematical Biology **54** (1992), no. 4, 521–536.

34. P. Trinh, A. McLysaght, and D. Sankoff, *Genomic features in the breakpoint regions between syntenic blocks*, Bioinformatics **20** (2004), 318–325.

35. E. Vidal and A. Marzal, *Fast computation of normalized edit distance*, IEEE Transactions on Pattern Analysis and Machine Intelligence **17** (1995), no. 9, 899–902.