Evolutionary Puzzles: An Introduction to Genome Rearrangement

Mathieu Blanchette

Department of Computer Science and Engineering
Box 352350
University of Washington
Seattle, WA 98195-2350 U.S.A.
206-543-5118
fax: 206-543-8331
blanchem@cs.washington.edu

Abstract. This paper is intended to serve as an introduction to genome rearrangement and its use for inferring phylogenetic trees. We begin with a brief description of the major players of the field (chromosomes, genes, etc.) and the types of mutations that can affect them, focussing obviously on genome rearrangement. This leads to a simple mathematical representation of the data (the order of the genes on the chromosomes), and the operations that modify it (inversions, transpositions, and translocations).
We then consider the problem of inferring phylogenetic (evolutionary) trees from genetic data. We briefly present the two major approaches to solve this problem. The first one, called distance matrix method, relies on the estimation of the evolutionary distance between each pair of species considered. In the context of gene order data, a useful measure of evolutionary distance is the minimum number of basic operations needed to transform the gene order of one species into that of another. This family of algorithmic problems has been extensively studied, and we review the major results in the field.
The second approach to inferring phylogenetic trees consists of finding a minimal Steiner tree in the space of the data considered, whose leaves are the species of interest. This approach leads to much harder algorithmic problems. The main results obtained here are based on a simple evolutionary metric, the number of breakpoints.
Throughout the paper, we report on various biological data analyses done using the different techniques discussed. We also point out some interesting open problems and current research directions.

1 Introduction

Understanding and classifying the incredible diversity of living beings has been the focus of much work, starting as early as in Ancient Greece. Since Darwin's thesis on evolution of species, we know that the diversity observed today is the result of a long process in which speciation (the event where two groups of

V.N. Alexandrov et al. (Eds.): ICCS 2001, LNCS 2074, pp. 1003–1011, 2001.

organisms from one species slowly diverge until they form two different, though closely related, species) played a key role. The history of these speciation events can be represented by a phylogenetic tree, where each leaf is labeled with a contemporary species and where the internal nodes correspond to hypothetical speciation events. Figure 1 illustrates the phylogenetic tree relating a group of vertebrates.

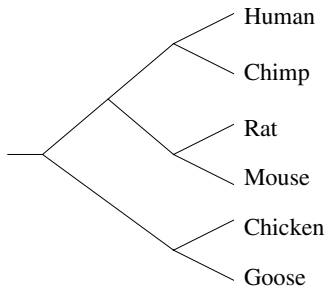

Fig. 1. The phylogenetic tree relating six species of vertebrates. Branch lengths are not to scale.

The problem of inferring the phylogenetic tree connecting a set of species is both of theoretical and practical importance. The problem has historically been addressed by considering morphologic, anatomic or developmental traits: species that have similar traits are likely to be closely related. However, for the last 30 years, most phylogenetic inference has been based on the DNA sequence of the species considered. The quantity of information contained in the genome of an organism is several orders of magnitude larger than that contained in its observable phenotype, and thus it potentially allows for a much more accurate inference. The DNA sequence of different organisms from different species differs because of mutations that occurred since their last common ancestor. It is by studying these mutations that one can hope to infer phylogenetic relationships.

This paper focuses on one specific type of mutations called genome rearrangement. In Section 2, we describe these mutations. In Section 3, we introduce the problem of inferring phylogenetic trees, and describe two classes of approaches that have been used.

2 Genome Rearrangement

The genome of an organism consists of a long string of DNA, cut into a small number of segments called chromosomes. Genes are stretches of the DNA sequence that are responsible for encoding proteins. Each gene has an orientation, either forward or backward, depending in which direction it is supposed to be

read. A chromosome can thus be abstracted as an ordered set of oriented genes. Most higher organisms' chromosomes are linear (their DNA sequence has a beginning and an end), but for lower organisms like bacteria, the chromosome is circular (their DNA sequence has no beginning or end).

The most common and most studied mutations operating on DNA sequences are local: they affect only a very small stretch on DNA sequence. These mutations include nucleotide substitutions (where one nucleotide is substituted for another), as well as nucleotide insertions and deletions. Most phylogenetic studies have been based on these types of mutations.

Genome rearrangement is a different class of mutation affecting very large stretches of DNA sequence. A genome rearrangement occurs when a chromosome breaks at two or more locations (called the breakpoints), and the pieces are reassembled, but in the "wrong" order. This results in a DNA sequence that has essentially the same features as the original sequence, except that the order of these features has been modified.

If the chromosome breaks occur in non-functional sequence, the rearrangement is unlikely to have any deleterious effects. On the other hand, a rearrangement whose breakpoints fall in functional sequence (e.g. genes) will almost certainly make the gene dysfunctional, rendering the offspring unlikely to survive. Consequently, almost all genome rearrangements that become fixed in future generations involve inter-genic breakpoints.

Figure 2 illustrates the three most common types of genome rearrangements. The first two, inversions and transpositions, affect only one chromosome at a time. The result of an inversion is to reverse the DNA segment between the two breakpoints. The order of the genes on this segment is reversed and their orientation is inverted. A transposition involves three breakpoints: the DNA segment between the two first breakpoints is spliced out and re-inserted somewhere else on the chromosome. A translocation involves two different chromosomes that exchange their ends.

The net effect of any genome rearrangement is to modify the order of features on a chromosome. The most important such features are genes, and most of the research done in this field has involved looking at genome rearrangement from the point of view of their effect on the gene order.

It is worth noticing that, compared to local mutations, genome rearrangements are extremely rare. However, over time, they accumulate and the order of the genes on each chromosome becomes more and more scrambled with respect to the original sequence. Thus, two closely related species will usually have similar gene orders (i.e. few genome rearrangements occurred during the evolution that separates them), whereas the gene orders of two more distant species will usually be less conserved.

3 Inferring Phylogenetic Trees

Inferring the phylogenetic (evolutionary) relationships among a set of species is a problem of both pure scientific and practical interest. This relationship is

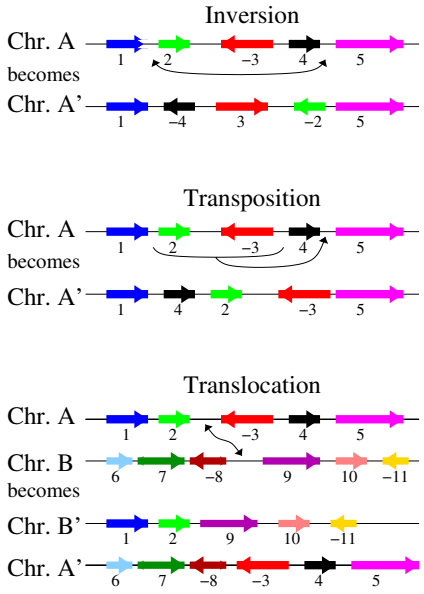

Fig. 2. Three most common types of genome rearrangements. A and B are chromosomes. Genes are numbered from 1 to 11.

usually depicted using a phylogenetic tree, whose leaves are labeled with the contemporary species under study, and whose internal structure indicates the order in which order these species diverged from each other. In some case, lengths will be associated with the branches of the tree. These lengths correspond to the amount of evolution that happened between the two endpoints of the branch.

By far the most common way to do phylogenetic inference is to study the evolution of DNA sequences from a local point of view. This approach has proven quite successful and has allowed us to infer the phylogenetic relationships among many species [14].

Gene arrangement can also be used to infer evolutionary relationships among species. Because genome rearrangements occur much more rarely than local mutations, it often allows to trace relationships between very distant species. Moreover, the fact that most genome rearrangements seem to have no effect at all on the fitness of the offspring makes our task easier. Consequently, one should see gene arrangement based studies as complimentary to sequence based studies.

3.1 Distance Matrix Methods

A first class of methods used to infer phylogenetic trees is based on the ability to estimate the amount of evolution that separates two species. Given species $S_1, S_2, ..., S_n$ with genomes $G_1, G_2, ..., G_n$, one can then compute the distance

matrix D, with D_{ij} = evolutionary distance between G_i and G_j. Notice that D can be calculated using any type of data: morphologic similarity, DNA sequence, gene arrangement, etc. Once the matrix D is computed, one finds the phylogenetic tree T and the length of each of its branches such that the distance between S_i and S_j on the tree T is as close as possible to that specified by D, according to some criterion (for example, the sum of square of errors). This optimization problem is NP-hard, but very good heuristics exist for trees with a small number of species [14].

But one question remains: how to evaluate the amount of evolution between two species? In our case, this translates to "how many rearrangements have happened between the order of the genes in G_i and that in G_j?"

It is obviously impossible to answer this question precisely: nobody was there when these rearrangements occurred, and some early rearrangement may have been undone later during evolution. Much research has been devoted to estimating rearrangement distance. The most popular distance measure between two gene orders is the edit-distance, defined as the smallest number of rearrangements needed to transform one gene order into the other. This measure is bound to be an underestimation of the true number of events that occurred, but in many cases it appears to be very close to the actual distance. In particular, as long as the true distance between two gene orders is not too large compared to the number of genes involved, the edit-distance will be very close to the true distance.

The problem of computing the edit-distance between strings has challenged computer scientists for a long time. Most of the research done regarding edit-distance between genomes has been done under the rather strong assumption that there is a known one-to-one correspondence between the genes of the two genomes compared. That is, no gene is duplicated and both genomes contain exactly the same set of genes (this can be obtained by ignoring some genes, if needed). When the genome is made of only one chromosome, the two gene orders can thus be assimilated to mathematical permutations, and the problem becomes to transform one permutation into the other. If the orientation of the genes is known, a sign (+ or -) is associated with each element of the permutation. When the genomes considered are circular, the associated permutations are also circular (i.e. the last and first elements of the permutation are adjacent).

The problem of computing edit-distances between gene orders was first studied by Watterson [15] and by Sankoff [9], who considered edit-distance restricted to inversions only. The most important result on computing edit-distance between permutations is a polynomial time algorithm to find the inversion distance between two signed permutations, due to Hannenhalli and Pevzner [7]. The algorithm can actually be generalized to find the minimal sequence of inversions and translocations between genomes containing more than one chromosome. An exact description of that algorithm is beyond the scope of this paper. Given the two permutations, the algorithm builds an edge-colored graph called the breakpoint graph. The minimal inversion distance is then given by the sum of four terms, each measuring some property of the breakpoint graph (the simplest term

is the number of connected components). The algorithm was later improved by [8] to run in time quadratic in the length of the sequences.

Notice that the knowledge of the orientation of each gene on the sequence is crucial to the time complexity of the algorithm. Indeed, the inversion distance problem for two unsigned permutations was shown to be NP-hard in [5]. No polynomial time algorithm for calculating the transposition distance is known, nor is any complexity result. A 3/2-approximation algorithm is described in [1].

Most algorithmic work has focussed on one type of rearrangement at a time, but nature doesn't have this restriction. In some groups of species, rearrangements occurring appear to be strongly biased towards one type of rearrangement, but most of the time, all three types of rearrangements can occur. Blanchette and Sankoff [4] have proposed a heuristic that computes the minimal-cost sequence of rearrangement between two gene orders, where each type of rearrangement is given a different cost. In nature, rearrangements involving short stretches of DNA seem to be more frequent than those involving longer segments. This could be taken into account when computing the minimal cost solution. Finally, it is believed that some inter-genic regions, called hot spots, are more prone to chromosome break than others. If this information was available, one could potentially use it to infer a more accurate evolutionary distance.

Edit-distance computation still offers many open and challenging problems. An important one is to generalize edit-distance computation to cases where genomes contain duplicated genes, and to the case where there is no clear one-to-one correspondence between the genes in the two genomes. The problem thus becomes an edit-distance problem on strings rather than on permutations.

Distance matrix methods have been used quite successfully to reconstruct phylogenetic trees. Many studies have used organelles (mitochondria and chloroplasts) genome to do their analysis. Organelles are small cellular structures that have their own genome, distinct from the nuclear genome. This genome is usually much smaller than the nuclear genome. For example, the mitochondrial genome of most metazoans have the same set of 37 genes. However, the order of these 37 genes varies a lot, and that makes them good candidates for phylogenetic inference (see, for example, [4] and [13]). Chloroplast genome have also been used for phylogenetic inference [6].

3.2 Reconstruction Methods

The distance matrix based inference methods are attractive because all they require is the ability to compute pairwise distances. However, this approach also has its downsides. First, the tree inferred doesn't contain any information about what the ancestral genomes looked like. In fact, it is likely that there is no evolutionary scenario that can match the tree and the branch lengths inferred. Second, the fact the tree is inferred from a $n \times n$ real number matrix means that much of the data is left behind. One could potentially do a better job by doing the tree inference directly using the data, without going first through a distance matrix.

This is exactly what reconstruction methods do. The problem is now framed as a Steiner tree problem: find the evolutionary scenario that most economically explains the genomes observed. More precisely:
Given: a set of genomes $G = \{g_1, ..., g_n\}$, each located in some metric space $< S, d >$ (for example, the space of all possible gene orders under some edit-distance metric)
Find: the set of ancestral genomes $A = \{a_1, ..., a_{n-2}\}$, where $a_i \in S \ \forall i$, and an unrooted tree $T = (G \cup A, E)$ with leaves G and internal nodes A, such that $\sum_{(v_i,v_j)\in E} d(v_i, v_j)$ is minimized.

This problem is NP-hard for most interesting metric spaces. The difficulty stems from two problems: i) there is a very large number of tree topologies, and ii) for each tree topology, there is a huge number of possible choices as to how locate the ancestral nodes. In fact, the very simplified Median problem, in which $n = 3$ and there is only one ancestral node, is NP-hard, even for signed inversions (for which the edit-distance computation is only of quadratic complexity). This leaves little hope to solve interesting problems with a larger number of species. However, good heuristics or approximation algorithms could be possible.

These rather depressing complexity results have motivated research towards finding new, simpler metric spaces in which the Steiner tree problem would be easier, while retaining as much biological relevance as possible. One such metric is the number of breakpoints between two permutations $A = a_1a_2...a_n$ and $B = b_1b_2...b_n$, which is defined as the smallest number of places where the chromosome A must be broken so that the pieces can be rearranged to form B. Notice that the pair of adjacent genes a_i, a_{i+1} in A needs to be broken only if a_i and a_{i+1} are not adjacent in B. When considering signed permutations, there is a breakpoint between a_i, a_{i+1} in A iff neither a_i follows a_{i+1} in B nor $-a_{i+1}$ follows $-a_i$ in B.

The number of breakpoints between two permutations can trivially be computed in time $O(n)$. This metric is not an edit-distance, but it has been shown to be closely related to the actual number of rearrangement events between two permutations [3]. Moreover, in many cases, different types of genome rearrangement occur with unknown probabilities, which makes the use of (weighted) edit-distance difficult. The breakpoint metric is not based on any specific type of rearrangement and thus can be applied without knowing anything about them.

Unfortunately, the Steiner tree problem under the number of breakpoint metric, and even the Median problem, is still NP-hard [11]. However, in this case, very good heuristics exist [12]. These heuristics rely on the ability to solve the Median problem by reducing it to a Traveling Saleman Problem (TSP), in which the number of cities is $2n$. TSP is itself an NP-hard problem, but it has been studied extensively, and very good heuristics have been developed [10]. This reduction allows one to solve quickly and optimally the Median problem for genomes containing less than a few hundred genes.

We then use our ability to solve the Median problem to iteratively assign near-optimal ancestral gene orders to the internal nodes of a fixed topology tree. Each tree topology is then evaluated in turn, and the topology requiring the

smallest number of breakpoints is chosen. Efficient programs were developed using this heuristic (BPAnalysis [12], [3], GRAPPA [6]). These programs have been used successfully for several phylogenetic inference based on mitochondrial and chloroplast genomes, respectively.

Many interesting problems associated with reconstruction methods remain to be considered. A few of them are outlined here. First, no good heuristic is known for the Median problem for any of the important edit-distance metrics. There is interesting theoretical and practical work to be done in that direction.

When considering the breakpoint metric, many issues remain. One of the most important of them is how to solve the Median problem when some of the input genomes have some genes missing, in which case the TSP reduction breaks down. This is an extremely important case, because with many data sets, the gene content of each genome is quite variable. The current approach is to consider only genes that occur in each of the input genomes, but that obviously throws away a lot of information that may be valuable for inferring the correct phylogenetic tree.

The problem of generalizing algorithms from permutations (where there is a one-to-one correspondence between the genes of two genomes) to strings (where each gene can occur several times in one genome) is also of great interest in the context of breakpoint distances. In fact, even the problem of computing the number of breakpoints between two strings (defined as the smallest number of times you need to cut string A to be able to rearrange the pieces to form string B), hasn't been solved yet.

4 Conclusion

Genome rearrangement is a great source of interesting problems for computer scientists. Inferring phylogenetic trees based on genome rearrangement often translates into nice, clean algorithmic problems. Many of these problems remain open. But genome rearrangements are not just puzzles for computer scientists. Most algorithms developed in this field have been applied to real biological data and have given good insights about the evolution of the species considered. With the various sequencing projects in progress, new data sets will become available. For example, the genomes of more than 30 bacteria and archebacteria have now been completely sequenced. These genomes contain more than 1000 genes each, and will to take genome rearrangement studies to a whole new scale. The whole genome of various higher organisms (worm, fruit fly, human, mouse, etc.) is also completely sequenced or close to being completed. These genomes contain a few tens to thousands of genes, and promise new and interesting algorithmic problems.

5 Acknowledgements

I will like to sincerely thank Saurabh Sinha for his help with the preparation of this manuscript.

References

1. V. Bafna and P.A. Pevzner. Sorting by transpositions. *Proceedings of the 6th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA 95),* 614-623, 1995.
2. V. Bafna and P.A. Pevzner. Sorting by reversals: Genome rearrangements in plant organelles and evolutionary history of X chromosome. *Molecular Biology and Evolution*, 12: 239-246, 1995.
3. M. Blanchette, T. Kunisawa and D. Sankoff. Gene order breakpoint evidence in animal mitochondrial phylogeny. *Journal of Molecular Evolution*, 49, 193-203, 1998.
4. M. Blanchette, T. Kunisawa and D. Sankoff. Parametric genome rearrangement. *Gene,* 172, GC:11-17, 1996.
5. A. Caprara. Sorting by Reversals is Difficult. *Proceedings of the First Annual International Conference on Computational Molecular Biology (RECOMB 97),* 75-83, 1997.
6. Cosner, M.E., Jansen, R.K., Moret, B.M.E., Raubeson, L.A., Wang, L.S., Warnow, T., and Wyman, S.. A new fast heuristic for computing the breakpoint phylogeny and a phylogenetic analysis of a group of highly rearranged chloroplast genomes.*Proc. 8th Int'l Conf. on Intelligent Systems for Molecular Biology ISMB-2000*, San Diego, 104-115, 2000.
7. S. Hannenhalli and P.A. Pevzner. Transforming men into mice (polynomial algorithm for genomic distance problem). *Proceedings of the IEEE 36th Annual Symposium on Foundations of Computer Science,* 581-592, 1995.
8. H. Kaplan, R. Shamir and R.E. Tarjan. Faster and Simpler Algorithm for Sorting Signed Permutations by Reversals. *Proceedings of the Eighth Annual ACM-SIAM Symposium on Discrete Algorithms (SODA 97),* 1997.
9. J. Kececioglu and D. Sankoff. Exact and approximation algorithms for sorting by reversals, with application to genome rearrangement. *Algorithmica,* 13: 180-210, 1995.
10. E.L. Lawler, J.K. Lenstra, A.H.G. Rinnooy Kan, and D.B. Shmoys The Travelling Salesman Problem. John Wiley and Sons, 1985.
11. I. Pe'er and R. Shamir. The median problems for breakpoints are NP-complete. *Electronic Colloquium on Computational Complexity, Technical Report* 98-071, 1998.
12. Sankoff, D. and Blanchette, M. Multiple genome rearrangement and breakpoint phylogeny. *Journal of Computational Biology* 5, 555-570, 1998.
13. D. Sankoff, G. Leduc, N. Antoine, B. Paquin, B.F. Lang and R. Cedergren. Gene order comparisons for phylogenetic inference: Evolution of the mitochondrial genome. *Proceedings of the National Academy of Sciences USA*, 89: 6575-6579, 1992.
14. Swofford, D.L., G.J. Olsen, P.J. Waddell, and D.M. Hillis. *In Molecular Systematics* (2nd ed., D.M. Hillis, C. Moritz, and B.K. Mable, eds.). Sinauer Assoc. Sunderland, MA. Ch. 11 (pp. 407-514), 1996.
15. G. A. Watterson, W. J. Ewens, T. E. Hall et A. Morgan. The chromosome inversion problem. *Journal of Theoretical Biology*, 99: 1-7, 1982.