Analysis of Genome Rearrangements

(submission for the Fourth DIAMCS International Implementation Challenge)

Sridhar Hannenhalli
Department of Computer Science and Engineering
The Pennsylvania State University
University Park, PA 16802

Note to Refree: Following text is directly taken from the paper "Transforming Men into Mice" by Sridhar Hannenhalli and Pavel Pevzner, to appear in FOCS-95. The paper is currently under revision and a copy will be available via anonymous ftp form ftp.cse.psu.eud as /pub/vision/hannenha/focs.ps latest by 10th august. However, a precursor to this work, by the same authors, entitled "Transforming Cabbage into Turnip" has appeared in STOC-95, pp 178-189. The previous work serves as the core in the solution for more general problem of genomic sorting. Although I plan to submit the code for the problem of computing reversal distance, I will talk about the general problem and discuss some improvements in the time complexity since the algorithm was first proposed.

Abstract: Many people (including ourselves) believe that transformations of humans into mice happen only in fairy tales. However, despite some differences in appearance and habits, men and mice are genetically very similar. In the pioneering paper, Nadeau and Taylor, 1984 estimated that surprisingly few genomic rearrangements (178 ± 39) happened since the divergence of human and mouse 80 million years ago. However, their analysis is non-constructive and no *rearrangement scenario* for man-mouse evolution has been suggested yet. The problem is complicated by the fact that rearrangements in multi-chromosomal genomes include inversions, translocations, fusions and fissions of chromosomes, a rather complex set of operations. As a result, at the first glance, a polynomial algorithm for the *genomic distance* problem with all these operations looks as improbable as the transformation of a (real) man into a (real) mouse. We prove a duality theorem which expresses the genomic distance in terms of easily computable parameters reflecting different combinatorial properties of sets of strings. This theorem leads to a polynomial-time algorithm for computing most parsimonious rearrangement scenarios. Based on this result and the latest comparative physical mapping data we have constructed a most parsimonious scenario of human-mouse evolution with 131 reversals/translocations/fusions/fissions. A combination of the genome rearrangement algorithm with the recently proposed experimental technique called ZOO-FISH suggests a new *constructive* approach to the 100-year old problem of reconstructing mammalian evolution.

Introduction:

Man-mouse comparative physical mapping started 20 years ago and currently more than 1300 pairs of homologous genes are mapped in these species. As a result, biologists found that the related genes in human and mouse are not chaotically distributed over the genomes but instead form "conserved blocks" (*synteny groups*).

Current comparative mapping data indicate that both human and mouse genomes are combined from approximately 150 blocks which are "shuffled" in human as compared to mouse (Copeland et al., 1993). Shuffling of blocks happens quite rarely (roughly once in a million years) thus making it possible to reconstruct a rearrangement scenario of human-mouse evolution. Below we present the combinatorial formulation of the problem.

In the model we consider, every *gene* is represented by an *identification* number (positive integer) and an associated *sign* ("+" or "−") reflecting the *direction* of the gene. A *chromosome* is defined as a *sequence* of genes, while a *genome* is defined as a *set* of chromosomes. Given two genomes Π and Γ with the same set of genes, we are interested in a most parsimonious scenario of *evolution* of Π into Γ, i.e. the shortest sequence of rearrangements (defined below) to transform Π into Γ. Fig. 1 illustrates 4 rearrangement events transforming one genome into another.

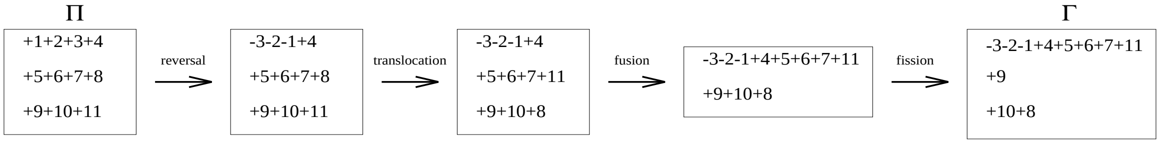

Figure 1: Evolution of genome Π into genome Γ

Let $\Pi = \{\pi(1), \ldots, \pi(N)\}$ be a genome consisting of N chromosomes and let $\pi(i) = (\pi(i)_1 \ldots \pi(i)_{n_i})$, n_i being the number of genes in the i^{th} chromosome. Every chromosome π can be viewed either from "left to right" (i.e. as $\pi = (\pi_1 \ldots \pi_n)$) or from "right to left" (i.e. as $-\pi = (-\pi_n \ldots -\pi_1)$) leading to two equivalent representations of the same chromosome. From this perspective a 3-chromosomal genome $\{\pi(1), \pi(2), \pi(3)\}$ is equivalent to $\{\pi(1), -\pi(2), \pi(3)\}$ or $\{-\pi(1), -\pi(2), -\pi(3)\}$, i.e. the *directions* of chromosomes are irrelevant. The four most common elementary rearrangement events in multi-chromosomal genomes are *reversals*, *translocations*, *fusions* and *fissions* as defined below.

Let $\pi = (\pi_1 \ldots \pi_n)$ be a chromosome and $1 \leq i \leq j \leq n$. A *reversal* $\rho(\pi, i, j)$ rearranges the genes *inside* the chromosome $\pi = (\pi_1 \ldots \pi_{i-1}\pi_i \ldots \pi_j\pi_{j+1} \ldots \pi_n)$ and transforms π into a chromosome $(\pi_1 \ldots \pi_{i-1} - \pi_j \ldots - \pi_i\pi_{j+1} \ldots \pi_n)$. Let $\pi = (\pi_1 \ldots \pi_n)$ and $\sigma = (\sigma_1 \ldots \sigma_m)$ be two chromosomes and $1 \leq i \leq n+1$, $1 \leq j \leq m+1$. A *translocation* $\rho(\pi, \sigma, i, j)$ exchanges genes *between* two chromosomes π and σ and transforms them into chromosomes $(\pi_1 \ldots \pi_{i-1}\sigma_j \ldots \sigma_m)$ and $(\sigma_1 \ldots \sigma_{j-1}\pi_i \ldots \pi_n)$ with $(i-1)+(m-j+1)$ and $(j-1)+(n-i+1)$ genes respectively. We denote $\Pi \cdot \rho$ as the genome obtained from Π as a result of a rearrangement ρ. Given genomes Π and Γ, the *genomic sorting* problem is to find a series of rearrangements (reversals and translocations) $\rho_1, \ldots, \rho_t$ such that $\Pi \cdot \rho_1 \cdots \rho_t = \Gamma$ and t is minimum. We call t the *genomic distance* between Π and Γ and denote it as $d(\Pi, \Gamma)$.

We distinguish between *internal* reversals which do not involve ends of the chromosomes (i.e. the reversals $\rho(\pi, i, j)$ of a n-gene chromosome with $1 < i \leq j < n$) and *prefix* reversals involving ends of the chromosomes (i.e. either $i = 1$ or $j = n$). Also note that a translocation $\rho(\pi, \sigma, n+1, 1)$ concatenates the chromosomes π and σ resulting in a chromosome $\pi_1 \ldots \pi_n\sigma_1 \ldots \sigma_m$ and an *empty* chromosome $\emptyset$. This special translocation leading to a reduction in the number of (non-empty) chromosomes is known in molecular biology as a *fusion*. The translocation $\rho(\pi, \emptyset, i, 1)$ for $1 < i \leq n$ "breaks" a chromosome π into two chromosomes $(\pi_1 \ldots \pi_{i-1})$ and $(\pi_i \ldots \pi_n)$. This translocation leading to an increase in the number of (non-empty) chromosomes is known as a *fission*. A translocation is *internal* if it is neither a fusion, nor a fission. Fusions and fissions are rather common in mammalian evolution. For example, the only difference in the overall genome organization of human and chimpanzee is the fusion of chimpanzee chromosomes 12 and 13 into human chromosome 2.

Genome rearrangements provide a multitude of challenges for computer scientists; see Pevzner and Waterman, 1995 for a review of combinatorial problems motivated by genome rearrangements. Kececioglu and Sankoff, 1993 suggested the first approximation algorithm to analyze genome rearrangements in uni-chromosomal genomes (reversals only). The problem was further studied by Bafna and Pevzner, 1993, Kececioglu and Sankoff, 1994, Kececioglu and Gusfield, 1994, Bafna and Pevzner, 1995 and Hannenhalli and Pevzner, 1995a. See also Sankoff et al.,1992, Bafna and Pevzner, 1995 and Hannenhalli et al., 1995 for biological applications, as well as Gates and Papadimitriou, 1979, Even and Goldreich, 1981, Jerrum, 1985, Aigner and West, 1987, Cohen and Blum, 1993, Heydari and Sudborough, 1993 for studies of related combinatorial problems.

Kececioglu and Ravi, 1995 made the first attempt to analyze rearrangements of multi-chromosomal genomes by devising an approximation algorithm for genomes evolving by internal reversals and internal translocations. Recently Hannenhalli, 1995 devised a polynomial algorithm for the case when only internal translocations are allowed. All these algorithms address the case when both genomes contain the same number of chromosomes. This is a serious limitation since different organisms (in particular human and mouse) have different numbers of chromosomes. From this perspective, every realistic model of genome rearrangements should include fusions and fissions. Moreover, despite some shortcomings (see the last section for a discussion on centromeres) the reversals/translocations/fusions/fissions model adequately reflects the existing biological challenges (Joe Nadeau, personal communication). It turns out that fusions and fissions present a major difficulty in analyzing genome rearrangements: the problem of devising an *approximation* algorithm for genome rearrangements with reversals/translocations/fusions/fissions was raised by Kececioglu and Ravi, 1995. This paper presents an *exact* polynomial algorithm for this problem.

Every analysis of rearrangements involves revealing "hidden obstacles" which prevent a "fast" transformation of one genome into another (like dual variables in linear programming). Sixty years ago Dobzhansky and Sturtevant, 1938 already used a notion of breakpoint, which is the most obvious example of such an obstacle. Based on this notion, Kececioglu and Sankoff, 1994, devised a 2-approximation algorithm for uni-chromosomal genomes. However, an estimate of genomic distance in terms of breakpoints is very inaccurate. Bafna and Pevzner, 1993 revealed another obstacle (cycle decomposition) which significantly improves the bounds for genomic distance. Finally, Hannenhalli and Pevzner, 1995a found a duality theorem for uni-chromosomal genomes (reversals only) which expresses the genomic distance in terms of four parameters describing the combinatorial structure of permutations. In case of multi-chromosomal genomes combinatorics of rearrangements becomes rather complicated. This paper presents the duality theorem for multi-chromosomal genomes which computes the genomic distance in terms of seven (!) parameters capturing different combinatorial properties of sets of strings.

Our analysis of multi-chromosomal genomes extensively uses the duality theorem for uni-chromosomal genomes (Hannenhalli and Pevzner, 1995a), Using a few notions of *flipping* of chromosomes and *capping* of chromosomes, we reduce the general problem of sorting multi-chromosomal genomes to the problem of sorting uni-chromosomal genomes.

References

[1] M. Aigner and D. B. West. Sorting by insertion of leading element. *Journal of Combinatorial Theory*, 45:306–309, 1987.

[2] V. Bafna and P. Pevzner. Genome rearrangements and sorting by reversals. In *34th Annual IEEE Symposium on Foundations of Computer Science*, pages 148–157, 1993. (to appear in SIAM J. Computing).

[3] V. Bafna and P. Pevzner. Sorting by reversals: Genome rearrangements in plant organelles and evolutionary history of X chromosome. *Molecular Biology and Evolution*, 12:239–246, 1995a.

[4] V. Bafna and P. Pevzner. Sorting by transpositions. In *Proc. 6th Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 614–623, 1995b.

[5] D. Cohen and M. Blum. Improved bounds for sorting pancakes under a conjecture. 1993 (manuscript).

[6] N. G. Copeland, N. A. Jenkins, D. J. Gilbert, J. T. Eppig, L. J. Maltals, J. C. Miller, W. F. Dietrich, A. Weaver, S. E. Lincoln, R. G. Steen, L. D. Steen, J. H. Nadeau, and E. S. Lander. A genetic linkage map of the mouse: Current applications and future prospects. *Science*, 262:57–65, 1993.

[7] T. Dobzhansky and A.H.Sturtevant. Inversions in the chromosomes of *drosophila pseudoobscura. Genetics*, 23:28–64, 1938.

[8] S. Even and O. Goldreich. The minimum-length generator sequence problem is NP-hard. *Journal of Algorithms*, 2:311–313, 1981.

[9] W. H. Gates and C. H. Papadimitriou. Bounds for sorting by prefix reversals. *Discrete Mathematics*, 27:47–57, 1979.

[10] S. Hannenhalli. Polynomial algorithm for computing translocation distance between genomes. In *Combinatorial Pattern Matching, Proc. 6th Annual Symposium (CPM'95)*, Lecture Notes in Computer Science, pages 162–176. Springer-Verlag, Berlin, 1995.

[11] S. Hannenhalli, C. Chappey, E. Koonin, and P. Pevzner. Scenarios for genome rearrangements: Herpesvirus evolution as a test case. In *Proc. 3rd Intl. Conference on Bioinformatics and Complex Genome Analysis*, pages 91–106, 1995.

[12] S. Hannenhalli and P. Pevzner. Transforming cabbage into turnip (polynomial algorithm for sorting signed permutations by reversals). In *Proc. 27th Annual ACM Symposium on the Theory of Computing*, pages 178–189, 1995a.

[13] M. Heydari and I. H. Sudborough. On sorting by prefix reversals and the diameter of pancake networks. 1993 (manuscript).

[14] M. Jerrum. The complexity of finding minimum-length generator sequences. *Theoretical Computer Science*, 36:265–289, 1985.

[15] J. Kececioglu and D. Gusfield. Reconstructing a history of recombinations from a set of sequences. In *5th Annual ACM-SIAM Symp. on Discrete Algorithms*, pages 471–480, 1994.

[16] J. Kececioglu and R. Ravi. Of mice and men: Evolutionary distances between genomes under translocation. In *Proc. 6th Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 604–613, 1995.

[17] J. Kececioglu and D. Sankoff. Exact and approximation algorithms for the inversion distance between two permutations. In *Combinatorial Pattern Matching, Proc. 4th Annual Symposium (CPM'93)*, volume 684 of *Lecture Notes in Computer Science*, pages 87–105. Springer-Verlag, Berlin, 1993. (Extended version has appeared in Algorithmica, 13: 180-210, 1995.).

[18] J. Kececioglu and D. Sankoff. Efficient bounds for oriented chromosome inversion distance. In *Combinatorial Pattern Matching, Proc. 5th Annual Symposium (CPM'94)*, volume 807 of *Lecture Notes in Computer Science 807*, pages 307–325. Springer-Verlag, Berlin, 1994.

[19] J. H. Nadeau and B. A. Taylor. Lengths of chromosomal segments conserved since divergence of man and mouse. *Proc. Natl. Acad. Sci. USA*, 81:814–818, 1984.

[20] P.A. Pevzner and M.S. Waterman. Open combinatorial problems in computational molecular biology. In *3rd Israel Symposium on Theory of Computing and Systems*, pages 158–163. IEEE Computer Society Press, 1995.

[21] 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. *Proc. Natl. Acad. Sci. USA*, 89:6575–6579, 1992.