November 21, 1994

DIMACS Implementation Challenge: Genome Rearrangements

Sequence alignment is often the first step in molecular evolution studies. However, in many cases sequence alignment is very unreliable thus making further evolutionary tree reconstruction almost impossible. As a result the classical methods of *sequence comparison* frequently lead to ambiguous conclusions for highly diverged genomes and alternative methods are sought. Since it is often found that the order of genes is much more conserved than the DNA sequence (Franklin, 1985) an approach based on comparison of *gene orders* (Sankoff et al.,1990,1992, Sankoff, 1993) versus traditional comparison of *DNA sequences* might be a method of choice for many "hard-to-analyze" genomes.

Similar problems exist for such conserved genomes as plant mitochondrial DNA (mtDNA). The point mutation rate in plant mtDNA is estimated to be 100 times slower than in animal mtDNA and many genes are $99\% - 99.9\%$ identical in related species (Palmer and Herbon, 1988). However, although there is little change in the DNA sequence of the plant mitochondrial genes from one species to another, there is rapid and extensive change in gene arrangements.

Analysis of genome rearrangements in molecular evolution and systematics was pioneered by Dobzhansky and Sturtevant, 1938, who published a mileston e paper with an evolutionary tree presenting a rearrangement scenario with 17 inversions for the species *D. pseudoobscura* and *D.miranda.* Studies of genome rearrangements frequently involve solving a combinatorial "puzzle" to find a shortest series of rearrangement to transform one genome into another. Fig. 1 presents the order of conserved blocks in

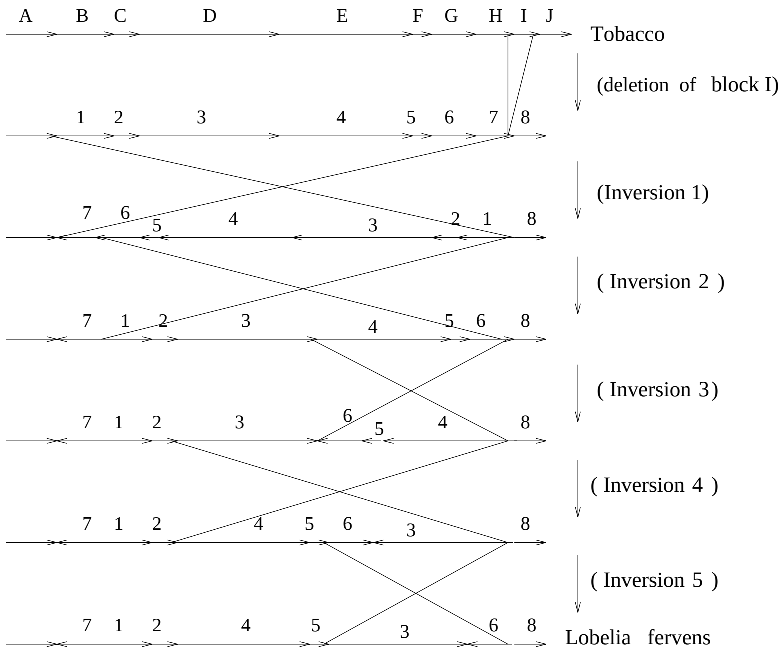
A B C D E F G H I J
Tobacco
(deletion of block I)
1 2 3 4 5 6 7 8
(Inversion 1)
7 6 5 4 3 2 1 8
(Inversion 2)
7 1 2 3 4 5 6 8
(Inversion 3)
7 1 2 3 6 5 4 8
(Inversion 4)
7 1 2 4 5 6 3 8
(Inversion 5)
7 1 2 4 5 3 6 8
Lobelia fervens

Figure 1:

Tobacco and *Lobelia fervens* chloroplast DNA (cpDNA) with a hypothetical sequence of rearrangement events (Knox et al.,1993) during evolution of *Lobelia fervens* from a tobacco-like ancestral genome. In the cases of genomes consisting of small number of "conserved blocks" biologists were able to find the most parsimonous scenarios for rearrangements. However, for genomes consisting of more than 10 blocks, exhaustive search over all potential solutions is far beyond the possibilities of "pen-and-pencil" methods.

With the advent of large-scale DNA mapping and sequencing, the number of biological problems similar to the one presented in Fig. 1 is rapidly growing in different areas, including evolution of plant cpDNA (Raubeson and Jansen, 1992, Hoot and Palmer, 1994) and mtDNA(Palmer and Herbon, 1988, Fauron and Havlik, 1989), animal mtDNA (Hoffman et al., 1992, Sankoff et al., 1992), virology (Koonin and Dolya, 1993, Hannenhalli et al., 1994), *Drosophila* genetics (Carson, 1983, Whiting et al, 1989) and comparative physical mapping (Lyon, 1988). *On the other hand, until recently there were no computer science results allowing a biologist to analyze genome rearrangements.* One of the goals of DIMACS Implementation Challenge 1995 is to fill this gap and to develop computer programs to analyze genome rearrangements.

Genomes evolve by inversions and transpositions as well as by deletion, insertion and duplication of fragments. Multi-chromosomal genomes also evolve by translocations. Inversions seem to be a very common rearrangement, in fact some genomes (for example, many plant mitochondrial DNA) are believed to rearrange almost solely by inversions (Palmer and Herbon, 1988) and less frequently by duplications. A combinatorial problem of *sorting by reversals* corresponding to genome rearrangements by inversions was introduced by Kececioglu and Sankoff, 1993 who suggested the first "performance guarantee" algorithm to analyze genome rearrangements.

Sorting by reversals corresponds to a simple model in which inversions of arbitrary length fragments are the only significant cause of evolutionary divergence. Despite some shortcomings, this model has been used by many biologists to analyze rearrangements in different mtDNAs and cpDNAs. Below a formal definition of sorting by reversals is given.

The order of genes in two organisms is represented by permutations $\pi = (\pi_1\pi_2\ldots\pi_n)$ and $\sigma = (\sigma_1\sigma_2\ldots\sigma_n)$. A *reversal* $\rho(i,j)$ of an interval $[i,j]$ is the transformation

$$\rho : \quad (\pi_1\ldots\pi_{i-1}\pi_{\mathbf{i}}\pi_{\mathbf{i+1}}\ldots\pi_{\mathbf{j-1}}\pi_{\mathbf{j}}\pi_{j+1}\ldots\pi_n) \rightarrow (\pi_1\ldots\pi_{i-1}\pi_{\mathbf{j}}\pi_{\mathbf{j-1}}\ldots\pi_{\mathbf{i+1}}\pi_{\mathbf{i}}\pi_{j+1}\ldots\pi_n)$$

which has the effect of reversing genes $\pi_{\mathbf{i}}, \pi_{\mathbf{i+1}}, \ldots \pi_{\mathbf{j-1}}, \pi_{\mathbf{j}}$. Given permutations π and σ, the *reversal distance problem* is to find a shortest series of reversals transforming genome π into genome σ, i.e. to find a series of reversals $\rho_1, \rho_2, \ldots, \rho_t$ such that $\pi \cdot \rho_1 \cdot \rho_2 \cdots \rho_t = \sigma$ and t is minimum. We call t the *reversal distance* between π and σ. Without loss of generality one can assume that σ is the *identity* permutation $\imath = (1, 2, \ldots, n)$ and in the following we consider a problem of transforming π into $\imath$. *Sorting π by reversals* is the problem of finding reversal distance $d(\pi)$, between π and $\imath$. As an example, Fig. 1, if considered from the bottom to the top presents sorting of *Lobelia fervens* permutation $\pi = 71245368$ by 5 reversals $\rho(6,7), \rho(4,7), \rho(5,7), \rho(2,7), \rho(1,7)$. Note that the bottommost reversal ρ on the interval $[6,7]$ reverses blocks 6 through 7 in π, i.e. $\rho : (71245\mathbf{36}8) \rightarrow (71245\mathbf{63}8)$.

Test Case

Recent advances in large-scale comparative genetic mapping offer exciting prospects for understanding mammalian genome evolution. The large number of conserved segments in the maps of man and mouse suggest that multiple chromosomal rearrangements have occured since the divergence of lineages leading to humans and mice. In their pioneering paper, Nadeau and Taylor, 1984 estimated that just 178 ± 39 rearrangements have occured since this divergence. This estimate survived a ten-fold increase in the amount of the comparative man/mouse mapping information; the new estimate based on the latest data (Copeland et al., 1993) changed little compared to Nadeau and Taylor, 1984.

Chromosomal rearrangements include pericentric and paracentric inversions, intrachromosomal and interchromosomal transpositions (see Bafna And Pevzner, 1995), reciprocal and Robertsonian translocations (see Kececioglu and Ravi, 1995), etc. (see Therman and Susman, 1 993 for a description of these rearrangements). A combinatorial analysis of all such rearrangements to derive a scenario of mammalian evolution represents a test case for DIMACS Implementation Challenge 1995. Other problems on genome rearrangements are listed below:

- sorting an unsigned permutation by reversals,
- sorting a signed permutation by reversals,
- sorting by adjacent block transpositions,
- sorting by transpositions and inversions so as to minimize the total number of operations,
- sorting by transpositions and inversions where the objective is to minimize a weighted sum of both operations

Inversions represent only one of several possible forms of rearrangements. Other common forms of rearrangements are transpositions, translocations, duplications, insertions and deletions of genome fragments. Moreover, these macroscopic changes actually coexist with point mutations that include insertions, deletions and substitutions of nucleotides. These problems provide a multitude of challenges, some much harder than others. Another computational challenge is to take into account "hot points" for rearrangements and to devise an algorithm which assigns different weights to different evolutionary events. Even more complicated problem is to analyze *multiple* genome comparisons (i.e. rearrangement scenarios involving many genomes) and to construct evolutionary trees describing genome rearrangements.

References

BAFNA V. and P. PEVZNER. 1993. Genome rearrangements and sorting by reversals. Proc. of 34th IEEE Symp. on Foundations of Computer Science. pp. 148–157.

BAFNA V.and P. PEVZNER. 1995. Sorting by transpositions. Proc. of 6th ACM-SIAM Symp. on Discrete Algorithms (to appear)

CARSON H. L. 1983. Chromosomal sequences and interisland colonization in Hawaiian Drosphila. Genetics. **103**:465–482.

COPELAND N. G., 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. 1993. A genetic linkage map of the mouse: Current applications and future prospects. Science. **262**:57–65.

DAVISSON M. 1987. X-linked genetic homologies between mouse and man. Genomics. **1**:213–227.

DOBZHANSKY T. and A.H.STURTEVANT. 1938. Inversions in the chromosomes of *drosophila pseudoobscura*. Genetics. **23**:28–64.

FAURON C.and M. HAVLIK. 1989. The maize mitochondrial genome of the normal type and the cytoplasmic male sterile type T have very different organization. Current Genetics. **15**:149–154.

FRANKLIN N. 1985. Conservation of genome form but not sequence in the transcription antitermination determinants of bacteriophages $\lambda, \phi 21$ and $P22$. Journal of

Molecular Evolution. **181**:75–84.

HANNENHALLI S., C. CHAPPEY, E. KOONIN and P. PEVZNER. 1994. Scenarios for genome rearrangements: Herpesvirus evolution as a test case. In Proc. of 3rd Intl. Conference on Bioinformatics and Complex Genome Analysis.

HOFFMANN R. J., J. L. BOORE and W. M. BROWN. 1992. A novel mitochondrial genome organization for the blue mussel, *Mytilus edulis*. Genetics. **131**:397–412.

HOOT S. B. and J. D. PALMER. 1994. Structural rearrangements including parallel inversions within the chloroplast genome of anemone and related genera. Journal of Molecular Evolution. **38**:274–281.

KECECIOGLU J. and D. SANKOFF. 1993. Exact and approximation algorithms for the inversion distance between two chromosomes. In Proc. of 4th Ann. Symp. on Combinatorial Pattern Matching, Lecture Notes in Computer Science **684**, pages 87–105. Springer Verlag. (To appear in Algorithmica, 1995 as "Exact and approximation algorithms for sorting by reversals, with application to genome rearrangements.")

KECECIOGLU J. and D. SANKOFF. 1994. Efficient bounds for oriented chromosome inversion distance. In Proc. of 5th Ann. Symp. on Combinatorial Pattern Matching. Proc. of 5th Ann. Symp. on Combinatorial Pattern Matching, Lecture Notes in Computer Science **807**, pages 307-325. Springer Verlag.

KECECIOGLU, J. and R. RAVI. 1995. Of mice and men: Algorithms for evolutionary distances between genomes with translocation. To appear in Proceedings of the 6th ACM-SIAM Symposium on Discrete Algorithms, January 1995.

KNOX E. B., S. R. DOWNIE and J. D. PALMER. 1993. Chloroplast genome rearrangements and evolution of giant lobelias from herbaceous ancestors. Mol. Biol. Evol. **10**:414–430.

KOONIN E. V. and V. V. DOLJA. 1993. Evolution and taxonomy of positive-strand RNA viruses: implications of comparative analysis of amino acid sequences. Crit.

Rev. Biochem. Molec. Biol. **28**:375–430.

Lyon M. F. 1988. X-Chromosome Inactivation and the Location and Expression of X-linked Genes. Am. J. Hum. Genet. **42**:008–016.

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

O'Brien S. J., J. E. Womack, L. A. Lyons, K. J. Moore, N. A. Jenkins and N. G. Copeland. 1993. Anchored reference loci for comparative genome mapping in mammals. Nature Genetics. **3**:103–112.

Palmer J. and L. Herbon. 1987. Unicircular structure of the Brassica hirta mitochondrial genome. Current Genetics. **11**:565–570.

Palmer J. D. and L. A. Herbon. 1988. Plant mitochondrial DNA evolves rapidly in structure, but slowly in sequence. Journal of Molecular Evolution. **27**:87–97.

Palmer J. D., B. Osorio and W. F. Thompson. 1988. Evolutionary significance of inversions in legume chloroplast DNAs. Current Genetics. **14**:65–74.

Raubeson L. A. and R. K. Jansen. 1992. Chloroplast DNA evidence on the ancient evolutionary split in vascular land plants. Science. **255**:1697–1699.

Sankoff D. 1993. Analytical approaches to genomic evolution. Biochimie. **75**:409–413.

Sankoff D., R. Cedergren and Y. Abel. 1990. Genomic divergence through gene rearrangement. pp. 428–438. In Molecular Evolution: Computer Analysis of Protein and Nucleic Acid Sequences, chapter 26. Academic Press, 1990.

Sankoff D., G. Leduc, N. Antoine, B. Paquin, B. F. Lang and R. Cedergren. 1992. Gene order comparisons for phylogenetic inference: Evolution of the mitochondrial genome. Proc. Natl. Acad. Sci. USA. **89**:6575–6579.

Sturtevant A. H. and T.Dobzhansky. 1936. Inversions in the third chromosome of wild races of *drosophila pseudoobscura*, and their use in the study of the history of the species. Proc. Nat. Acad. Sci. U.S.A. **22**:448–450.

Therman E. and M. Susman. 1993. Human Chromosomes, structure, behavior, and effects. Springer-Verlag.

Watterson G. A., W. J. Ewens, T. E. Hall and A. Morgan. 1982. The chromosome inversion problem. Journal of Theoretical Biology. **99**:1–7.

Whiting J., M. Pliley, J. Farmer and D. Jeffery. 1989. In situ hybridization analysis of chromosomal homologies in *Drosophila melanogaster* and *Drosophila virilis*. Genetics. **122**:99–109.