Reversals of Fortune

David Sankoff[1], Chungfang Zheng[2], and Aleksander Lenert[3]

[1] Department of Mathematics and Statistics
[2] Department of Biology
[3] Program in Biopharmaceutical Science,
University of Ottawa, Canada
{sankoff, czhen033, alene096}@uottawa.ca

Abstract. The objective function of the genome rearrangement problems allows the integration of other genome-level problems so that they may be solved simultaneously. Three examples, all of which are hard: 1) Orientation assignment for unsigned genomes. 2) Ortholog identification in the presence of multiple copies of genes. 3) Linearisation of partially ordered genomes. The comparison of traditional genetic maps by rearrangement algorithms poses all these problems. We combine heuristics for the first two problems with an exact algorithm for the third to solve a moderate-sized instance comparing maps of cereal genomes.

1 Introduction

The first chromosomal map dates from 1913 [30], at the same time the definitive chromosomal theory of heredity [19] was being elaborated. Soon comparative mapping had become an integral part of genetic research, e.g., Fig. 1 in [31], published in 1921. Long before the genomic era, comparative maps existed for *Drosophila* and other insects, mammals, including humans, livestock and rodents, cereals and other cultivars and other eukaryotic and prokaryotic groups.

Despite their immediate availability and the wealth of evidence they contain about evolutionary history, traditional comparative maps were bypassed when genome rearrangement algorithms ([14,15]), inspired by analyses of organelle and other small genomes (e.g., [23,29]), were adapted for direct use on DNA segments derived from whole nuclear genome sequences [24,2,4].

In this paper we discuss an approach to the application of rearrangement methods to traditional comparative maps, i.e., maps based on estimates of gene and marker locations in nuclear genome, and not directly on genome sequence. First, what are the difficulties we encounter when we attempt this?

Coarseness. Lack of resolution of the maps, i.e., two or more genes being mapped to the same position in one of the genomes. Genome rearrangement algorithms require that the input markers be totally ordered along each chromosome.

Missing Data. Order ambiguity in composite maps. Because maps constructed from a single type of experimental data usually contain a limited number of markers, we are motivated to combine maps for the same genome from

A. McLysaght et al. (Eds.): RECOMB 2005 Ws on Comparative Genomics, LNBI 3678, pp. 131–141, 2005.

different sources. Two genes or markers which are not ordered by any of the component maps will often remain unordered in the composite map. Again, rearrangement algorithms require that the input markers be totally ordered along each chromosome.

No Signs. No information about reading direction, i.e., which DNA strand the gene or marker is on. This information is not available from many of the methods used to construct maps. Genome rearrangement algorithms require this "orientation" information for efficient and exact execution.

Matches. Uncertain orthology.

Notation. Different nomenclatural traditions in the genetics communities producing the chromosomal maps for two species mean different annotations and difficulties for the analyst in deciding which markers in one genome correspond to markers in the other. Rearrangement algorithms require that genes or other markers on the two genomes be unequivocally paired as being derived from a single entity in an ancestral genome.

Paralogy. Several copies or near copies of a gene in a map. This leads to a one-to-many or many-to-may correspondence between the two maps. Genome rearrangement algorithms require one-to-one correspondences as input.

Conflicts. When two or more relatively sparse maps of a genome, compiled from different sources, are combined prior to comparison with the map of another genomes, there is often conflict concerning the orders of a some of the markers on both maps.

With the possible exception of **paralogy**, these difficulties are neatly avoided when complete genome sequences are being compared at the sequence level [24,2,4], though of course there are many other technical problems to be solved in that approach.

The difficulties listed above all have in common that we are missing some essential biological information required to carry out genome rearrangement analysis. Moreover, in each case (except **notation**) THE GENOME REARRANGEMENT PROBLEM MAY BE REFORMULATED IN SUCH A WAY THAT THE SOLUTION NOT ONLY PROVIDES A MINIMAL SERIES OF REVERSALS AND/OR TRANSLOCATIONS NECESSARY TO TRANSFORM ONE GENOME INTO ANOTHER, BUT ALSO SUPPLIES AN OPTIMAL ESTIMATE OF THE MISSING INFORMATION. It is the comparative context, together with the rearrangement-minimizing objective function, which "fills in" the gaps in our biological knowledge in the most reasonable way. This unexpected bounty from the rearrangement analysis is what is alluded to in the title of this paper.

Exact algorithms have been published to take care of **coarseness**, **missing data**, **no sign** and **paralogy**, all requiring exponential worst-case computing time. The latter two, the topics of Sections 3 and 4, respectively, have been proved NP-hard, and we have conjectured as much for the first two, which are the main focus of this paper, as discussed in Section 5. As for the **notation** problem, we may rely on one of the curated comparative browsers, such as Gramene [36] for cereals and some other plants, the NCBI Human-Mouse homology maps [20], UCSC Genome browser [35] for animals, or CompLDB [21] for livestock.

Solution of a typical comparative map rearrangement problem would require treating at least **coarseness**, **no sign** and **paralogy** simultaneously, and usually **missing data** and **conflict** as well. We will state the pertinent combinatorial optimization problem, but its exact solution would be feasible only on very particular, small instances. We do, however, give results of applying an exact algorithm allowing for **coarseness** and **missing data**, in all generality, applied to data where **no sign**, **paralogy** and **conflict** are dealt with heuristically during preprocessing, using some of the key ideas in their respective algorithms.

In Section 2, however, we will start with the essentials of genome rearrangement theory.

2 The Bicoloured Graph in Rearrangement Algorithms

Hannenhalli and Pevzner [15] showed how to find a shortest sequence of reversals and translocations that transform one completely specified genome χ with n genes on k chromosomes into another genome ψ of the same size but with h chromosomes, in polynomial time. Completely specified means that each chromosome is totally ordered, the sign of each gene is known, and there is no paralogy.

As described in [34], we construct a bicoloured graph on $2n + 2k$ vertices that decomposes uniquely into a set of alternating-coloured cycles and $h + k$ alternating-colour paths. First, each gene x in χ determines two vertices, x_t and x_h. Two dummy vertices e_{i_1} and e_{i_2} are added to the ends of each chromosome χ_i. The adjacencies in χ determine red edges. If x is the left neighbour of y in χ, and both have positive polarity, then x_h is connected by a red edge to y_t. If they both are negative, x_t is joined to y_h. If x is positive and y negative, or x is negative and y positive, x_h is joined to y_h, or x_t is joined to y_t, respectively. If x is the first gene in χ_i, then e_{i_1} is joined to x_t or x_h depending on whether x has positive or negative polarity, respectively. If x is the last gene, then e_{i_2} is joined to x_t or x_h depending on whether x is negative or positive.

Black edges are added according to the same rules, based on the adjacencies in genome ψ, though no dummy vertices are added in this genome.

Each vertex is incident to exactly one red and one black edge edge, except for the dummies in χ and the (non-dummy) vertices at the ends of chromosomes in ψ, which are each incident to only a red edge. The bicoloured graph decomposes uniquely into a number of alternating cycles plus $h + k$ alternating paths terminating in either the dummy vertices of χ or the end vertices of ψ, or one of each. Suppose the number of these paths that terminate in at least one dummy vertex is $j \leq h + k$. If the number of cycles is c, then the minimum number of reversals r and translocations t necessary to convert χ into ψ is given by:

$$r + t = n - j - c + \theta \tag{1}$$

where θ is a correction term that is usually zero for simulated or empirical data. For simplicity of exposition, we ignore this correction here. Indeed, in a recent framework [11] allowing p transpositions and more general block interchanges via circular intermediate chromosomal fragments, $\theta \equiv 0$, and we simply have

$$r + t + 2p = n - j - c. \tag{2}$$

The $n - j - c$ actual rearrangement steps for transforming χ into ψ can then be found via certain well-defined operations on the cycles of the bicoloured graph.

3 Sign Assignment

Our first problem is that of adding signs to an unsigned genome so as to a achieve a minimal reversal distance to the identity permutation $1, \cdots, n$. This is equivalent to the problem of sorting an unsigned permutation, known to be NP-hard [7].

As conjectured in [17] and proved in [16], for all segments of the permutation consisting of three or more consecutive integers (strips) in increasing order, plus signs can be given to all these integers, and for all decreasing strips, minus signs can be given, and this assignment is consistent. with a solution. In [16], it is also shown how to give signs to 2-strips. The algorithm these authors develop is exponential only in s, the number of singletons, and is polynomial if s is $O(\log n)$. Unfortunately, in comparative maps s often seems closer to $O(n)$.

Though there is much recent literature on approximation to unsigned reversal distance, relatively little work has been done on exact algorithms. Caprara *et al.* [8] have implemented a branch-and-price algorithm that enables the rapid sorting of up to 200 elements. Tesler (personal communication) has extended the approach in [16] to reversal and translocation distance, and implemented it in GRIMM [33].

4 Duplicate Genes, Paralogy, Gene Families

When there are paralogs, gene orders cannot be modeled as permutations, but only as more general strings. Though sorting strings by reversals can be done in polynomial time, this does not automatically give the reversal distance between strings, in contrast to sorting permutations by reversals, which is equivalent to calculating reversal distance. Indeed, reversal distance for strings is NP-hard [26].

The problem in analysing genomes containing paralogs is how to decide which paralog in one genome should be identified with which one in the other genome, in a biologically meaningful way. Thus string-based analyses that attempt to match all or as many as possible of the paralogs of a gene in one genome to distinct paralogs in the other are only meaningful under the often questionable assumption that all paralogs were present in the common ancestor genome.

A less ambitious, but biologically more reasonable, approach is to try to match only one paralog of each gene in one genome to one in the other, such that the gene orders of the matched paralogs (the *exemplars*) of each family result in a minimal reversal distance [27]. The hypothesis motivating this is that genomes reduced to contain only these exemplars will better tend to reflect the actual reversal history than reduced genomes made up of any other choice of exemplars, using a parsimony argument.

There is a growing literature on the problem of incorporating paralogy into genome rearrangement theory. This is most meaningfully carried out within the phylogenetic context [28,3], taking into account that the origin of paralogs in duplication events may occur on earlier or later branches of the evolutionary tree. In addition to work characterizing, approximating or generalizing the exemplar approach [6,22], there is research on rearrangement in the context of string theory [10,26], conserved interval/block theory [1,3] and other a number of other approaches [9,32]. Virtually all of these are based on the same principle, that matching of paralogs should minimize the rearrangement distance

5 Partial Order

A linear map of a chromosome that has several genes or markers at the same position π, because their order has not been resolved, can be reformulated as a partial order, where all the genes before π are ordered before all the genes at π and all the genes at π are ordered before all the genes following π, but the genes at π are not ordered amongst themselves.

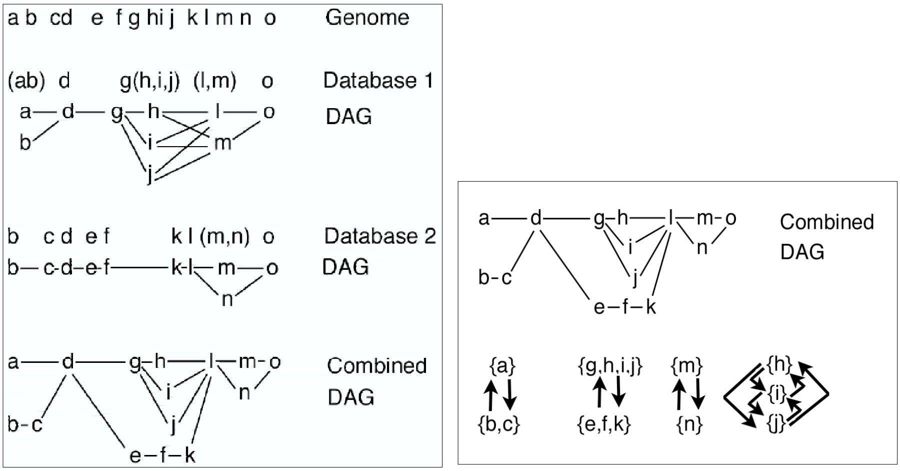

Fig. 1. (Left) Construction of DAGs from individual databases each containing partial information on genome, due to missing genes and missing order information, followed by construction of combined DAG representing all known information on the genome. All edges directed from left to right. (Right) Edges added to DAG to obtain DG containing all linearization as paths (though not all paths in the DG are linearizations of the DAG!). Each arrow represents a set of directed edges, one from each element in one set to each element of the other set.

For genomes with two or more gene maps constructed from different kinds of data or using different methodologies, there is only one meaningful way of combining the order information on two (partially ordered) maps of the same chromosome containing different subsets of genes. Assuming there are no conflicting order relations ($a < b, b < a$) nor conflicting assignments of genes to chromosomes among the data sets, for each chromosome we simply take the union of the partial orders, and extend this set through transitivity.

All the partial order data on a chromosome can be represented in a directed acyclic graph (DAG) whose vertex set is the union of all gene sets on that chromosome in the contributing data sets, and whose edges correspond to just those order relations that cannot be derived from other order relations by transitivity. The outcome of this construction is illustrated on the left of Figure 1.

We can extend genome rearrangement theory to the more general context where all the chromosomes are general DAGs rather than total orders [37,38]. The rearrangement problem becomes: to infer a transformation sequence (translocations and/or reversals) for transforming a set of linearizations (topological sorts), one for each chromosomal DAG in the genome of one species, to a set of linearizations of the chromosomal DAGs in the genome of another species, minimizing the number of translocations and reversals required.

A DAG can generally be linearized in many different ways, all derivable from a topological sorting routine. All the possible adjacencies in these linear sorts can be represented by the edges of a directed graph (DG) containing all the edges of the DAG plus two edges of opposite directions between all pairs of vertices, which are not ordered by the DAG. This is illustrated on the right in Figure 1.

We can make a bicoloured graph from the set of edges in the DGs for two partially ordered genomes. In the resulting graph, each of the DAG edges and both of the edges connecting each of the unordered pairs in the DG for each chromosome represent potential adjacencies in our eventual linearization of a genome. The n genes or markers and $2k$ dummies determine $2n + 2k$ vertices and the potential adjacencies determine the red and black edges, based on the polarity of the genes or markers. Where the construction for the totally ordered genomes contains exactly $n + k$ red edges and $n - h$ black edges, in our construction in the presence of uncertainty there are more potential edges of each colour, but only $2n + k - h$ can be chosen in our construction of the cycle graph, which is equivalent to the simultaneous linearization by topological sorting of each chromosome in each genome. It is this problem of selecting the right subset of edges that makes the problem difficult (and, we conjecture, NP-hard.)

Our approach to this problem is a depth-first branch and bound search in the environment of $h + k$ continually updated partial orders, one for each chromosome in each genome. The strategy is to build cycles and paths one at a time. After each one is completed, the current best construction serves as a bound to compare against the maximum number of cycles and paths that could possibly be built with the remaining eligible edges. The effect of the current bound becomes greater every time a potential edge is chosen for the graph, because this generally makes many other edges ineligible to be chosen at later steps. This is not just a question of avoiding multiple edges of the same colour incident to a single vertex, but also combinations of edges that are incompatible with one of the DAGs.

We have focused here on obtaining the cycle decomposition; this is equivalent to optimally linearizing the partial orders, so that finding the rearrangements themselves can be done using the previously available algorithms and software, e.g., GRIMM [33].

One problem we have not dealt with is **conflict**; different maps of the same genome do occasionally conflict, either because $b < a$ in one data set while $a < b$ in the other or because a gene is assigned to different chromosomes in the two data sets. There are a variety of possible ways of resolving order conflicts or, equivalently, of avoiding any cycles in the construction of the DAG. One way is to delete all order relations that conflict with at least one other order relation. Another is to delete a minimal set of order relations so that all conflicts can be resolved. Perhaps the approach that best balances loss of information with ease of application and interpretation is to discard a minimum set of gene occurrences so that all order conflicts are resolved. This method also resolves conflicts due to gene assignment to different chromosomes. Any gene that is discarded from all the data sets for one genome has, of course, to be discarded from the other.

6 Synthesis and Application

Given a map comparison that suffers from some combination of **coarseness**, **missing data**, **no sign** and **paralogy**, we can ask: simultaneously find the exemplars and sign assignments resulting in a minimum number of translocations and inversions necessary to transform some DAG linearisation of one genome into some DAG linearisation of the other. Since all three component problems are hard, there is scant hope that their combination is tractable. In this section, we describe a practical approach to one problem of this type.

Note that if there is **conflict**, we might want to avoid discarding exemplars in resolving conflict; if that is impossible, then we should at least take into account the sizes of any discarded gene families in assuring a minimum of genes occurrences are discarded. In any case, this minimum should be established beforehand, and should constrain the exemplar selection, if this is an issue. Under this one constraint, the goal is the minimization of genomic distance over all combinations of exemplar choices, eligible conflict resolutions, sign assignments and DAG linearisations,

The particular application we study, using the implementation of the DAG linearisation described in [38], is the comparison of the maize and sorghum genomes. We used one set of genomic markers for maize [25] and two for sorghum [18,5] as accessible in Gramene [36]. We extracted all markers registered as having homologs in maize and at least one of the sorghum datasets. This gave 463 marker occurrences in maize and 387 in sorghum, based on 296 total non-homologous markers.

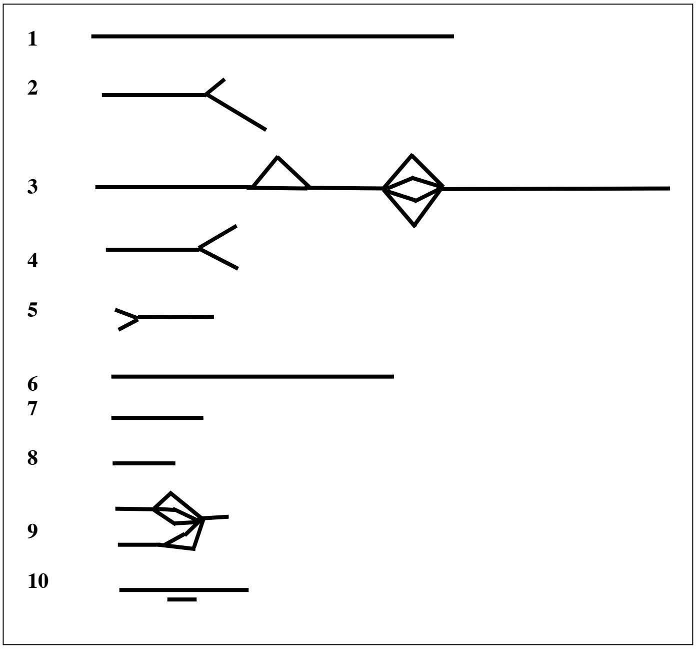

Fig. 2. DAGs for 10 sorghum chromosomes, scaled by number of markers analysed

Partly because this size of this problem is excessive for our implementation, we ignored any pair of chromosomes, one in maize and one in sorghum, with less than four markers in common. Some threshold, though perhaps not as large as four, is also justified by the facts that occasional syntenies of this sort are often the result of marker homology assignment or other error, and that especially in the case of singletons, the rearrangement solution simply includes two or three rearrangements solely to account for the position of this marker, and is independent of the rearrangement of the rest of the genome. This step left us with 381 marker occurrences in maize and 301 in sorghum, based on 263 total non-homologous markers. Thus by removing only 11% of the non-homologous markers from the original data, we remove 65 % of the excess paralogs, consistent with our suspicion that these do not represent orthologies.

As a next step, we identified all strips, as this is crucial not only to solving **no signs**, but is also helpful for **paralogy** and **conflict**. To take further advantage of strips, we removed paralogs and markers involved conflicts whenever they interrupted contiguous strips. We then found the exemplars for the remaining paralogies and resolved the remaining conflicts. To further reduce the size of the problem, we discarded a number of other singletons.

The remaining markers in the two sorghum and one maize datasets, representing 191 different markers, organized into 99 strips and singletons, could then be input into our exact linearisation algorithm. The DAGs for the sorghum chromosomes are illustrated in Figure 2. The solution involved 6 non-trivial

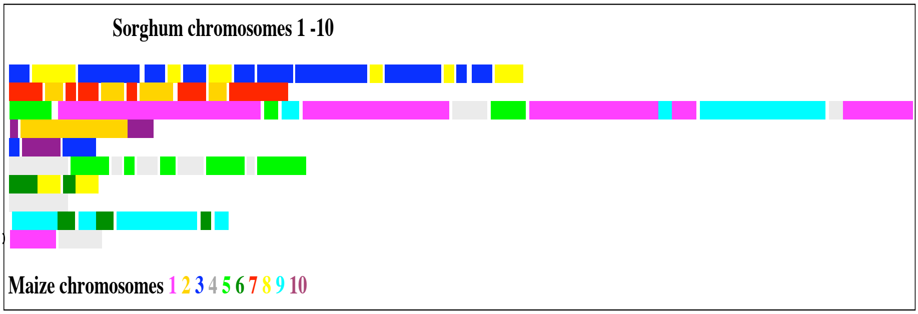

Fig. 3. Conserved segments on sorghum chromosomes, scaled by number of markers analysed

cycles (more than two edges) and 20 paths, implying a total of 73 inversions and translocations. Figure 3 portrays the configuration of the conserved segments in the two genomes, disposed on the sorghum chromosomes.

7 Discussion

A generally usable algorithm for the simultaneous solution of the linearisation and sign assignment problems seems feasible, since both can be handled within the partial order framework, though of course this is still a worst-case hard problem. There are many approaches possible to improve the current bound, to find a better sequence of edges as candidates to add to the current alternating colour cycle, and to incorporate heuristics, such as formalizing our strip-maximization/singleton-minimization procedure for discarding the most likely erroneous markers.

The situation with paralogy and conflict is more complicated, as the strong constraint of acyclicity in the DAG representation of the map data cannot be satisfied. Nevertheless, there is hope for some method drawn from the homology assignment literature we have cited to be incorporated into the solution of these problems in the comparative map context. The maize genome is known to have originated in a genome doubling event [13]; thus the treatment of duplicates through the exemplar or similar paradigm may be less appropriate than a genome halving analysis [12], which is only of polynomial complexity.

Acknowledgements

Research supported in part by grants from the Natural Sciences and Engineering Research Council of Canada (NSERC). DS holds the Canada Research Chair in Mathematical Genomics and is a Fellow of the Evolutionary Biology Program of the Canadian Institute for Advanced Research. Thanks to Glenn Tesler for a discussion of strategies in his GRIMM software.

References

1. Blin, G. and Rizzi, R. 2005. Conserved interval distance computation between non-trivial genomes. In Wang, L. (ed), Proceedings of COCOON ‘05. LNCS 3595. Berlin, Heidelberg:Springer Verlag, in press.
2. Bourque, G., Pevzner, P.A. and Tesler, G. 2004. Reconstructing the genomic architecture of ancestral mammals: lessons from human, mouse, and rat genomes. Genome Research, 14:507–516.
3. Bourque, G., Yacef, Y., and El-Mabrouk, N. 2005. Maximizing synteny blocks to identify ancestral homologs. manuscript.
4. Bourque, G., Zdobnov, E., Bork, P., Pevzner, P. and Tesler, G. 2005. Comparative architectures of mammalian and chicken genomes reveal highly variable rates of genomic rearrangements across different lineages. Genome Research, 15:98–110.
5. Bowers, J. E., Abbey, C., Anderson, S., Chang, C., Draye, X., Hoppe, A. H., Jessup, R., Lemke, C., Lennington, J., Li, Z., Lin, Y. R., Liu, S. C., Luo, L., Marler, B. S., Ming, R., Mitchell, S.E., Qiang, D., Reischmann, K., Schulze, S. R., Skinner, D. N., Wang, Y. W., Kresovich, S., Schertz, K. F., Paterson, A. H. 2003. A high-density genetic recombination map of sequence-tagged sites for sorghum, as a framework for comparative structural and evolutionary genomics of tropical grains and grasses. Genetics 165:367–86.
6. Bryant, D. 2000. The complexity of calculating exemplar distances. In Sankoff, D. and Nadeau, J. (eds), Comparative Genomics. Dordrecht, NL: Kluwer, pp. 207–212.
7. Caprara, A. 1997. Sorting by reversals is difficult. in Istrail, S., Pevzner, P.A. and Waterman, M.S. (eds.) Proceedings of the First Annual International Conference on Computational Molecular Biology (RECOMB’97), ACM Press, pp. 75–83.
8. Caprara, A., Lancia, G. and Ng, S.K. 2001. Sorting permutations by reversals through branch-and-price. INFORMS Journal on Computing 13, 224–244.
9. Chen, X., Zheng, J., Fu, Z., Nan, P., Zhong, Y., Lonardi, S. and Jiang, T. 2005. Assignment of orthologous genes via genome rearrangement. IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB), in press.
10. Christie, D. A. 1996. Sorting permutations by block interchanges. Information Processing Letters 60:165–169.
11. Friedberg, R., Attie, O. and Yancopoulos, S. 2005. Efficient sorting of genomic permutations by translocation, inversion and block interchange. Bioinformatics, in press
12. l-Mabrouk, N. and Sankoff, D. 2003. The reconstruction of doubled genomes. SIAM Journal on Computing 32:754–792.
13. Gaut, B. S. and Doebley, J.F. 1997. DNA sequence evidence for the segmental allotetraploid origin of maize. Proc. Natl. Acad. Sci. U S A. 94:6809–6814.
14. Hannenhalli, S. and Pevzner, P.A. 1995. Transforming cabbage into turnip (polynomial algorithm for sorting signed permutations by reversals). In Proc. 27th Annual ACM Symposium on the Theory of Computing, pp. 178–189.
15. Hannenhalli, S. and Pevzner, P.A. 1995. Transforming men into mice (polynomial algorithm for genomic distance problem. Proceedings of the IEEE 36th Annual Symposium on Foundations of Computer Science. 581–92.
16. Hannenhalli, S. and Pevzner, P.A. 1996. To cut or not to cut (applications of comparative physical maps in molecular evolution). In Proceedings of the 7th annual ACM-SIAM symposium on discrete algorithms, Philadelphia: SIAM, pp. 304–313.
17. Kececioglu, J. and Sankoff, D. 1993. Exact and approximation algorithms for the inversion distance between two permutations. In Proceedings of 4th Combinatorial Pattern Matching symposium, LNCS 684, Springer Verlag, pp. 87–105. (Cf. Algorithmica 13: 180–210, 1995.).

18. Menz, M. A., Klein, R. R., Mullet, J. E., Obert, J. A., Unruh, N.C., and Klein, P. E. 2002. A high-density genetic map of Sorghum bicolor (L.) Moench based on 2926 AFLP, RFLP and SSR markers. Plant Molecular Biology 48:483–99.
19. Morgan, T. H., Sturtevant, A. H. , Muller, H . J., and Bridges, C.B. 1915. The mechanism of Mendelian heredity. New York: Henry Holt. and Co.
20. NCBI Human Mouse Homology.http://www.ncbi.nlm.nih.gov/Homology/
21. Nicholas, F.W., Barendse, W., Collins, A., Darymple, B.P., Edwards, J.H., Gregory, S., Hobbs, M., Khatkar, M.S., Liao, W., Maddox, J.F., Raadsma, H.W. and Zenger K. R. 2004. Integrated maps and Oxford grids: maximising the power of comparative mapping. Poster at International Society of Animal Genetics.
22. Nguyen, C.T., Tay, Y.C. and Zhang, L. 2005. Divide-and-conquer approach for the exemplar breakpoint distance. Bioinformatics, in press.
23. Palmer, J.D., Osorio, B. and Thompson, W.F. 1988. Evolutionary significance of inversions in legume chloroplast DNAs. Current Genetics 14:6574.
24. Pevzner, P.A. and Tesler, G. 2003. Human and mouse genomic sequences reveal extensive breakpoint reuse in mammalian evolution. Proc Natl Acad Sci USA 100:7672–7
25. Polacco, M.L.; Coe, E., Jr. 2002. IBM Neighbors: A Consensus Genetic Map. (http://www.maizegdb.org/ancillary/IBMneighbors.html)
26. Radcliffe, A.J., Scott, A.D. and Wilmer, R.E. 2005. Reversals and transpositions over finite alphabets. SIAM Journal on Discrete Math, in press.
27. Sankoff, D. 1999. Genome rearrangement with gene families. Bioinformatics, 15:909–917.
28. Sankoff, D. and El-Mabrouk, N. 2000. Duplication, rearrangement and reconciliation. In Sankoff, D. and Nadeau, J. H., (eds) Comparative Genomics. Dordrecht, NL: Kluwer.
29. Sankoff, D., Leduc, G., Antoine, N., Paquin, B. Lang, B.F. and Cedergren, R. 1992. Gene order comparisons for phylogenetic inference: Evolution of the mitochondrial genome. Proc. Natl. Acad. Sci. USA, 89:6575–6579.
30. Sturtevant, A. H. 1913 The linear arrangement of six sex-linked factors in Drosophila, as shown by their mode of association. Jour. Exp. Zool. 14: 43–59.
31. Sturtevant, A. H. 1921. Genetic studies on *Drosophila simulans.* II. Sex-linked group of genes. Genetics 6: 43–64.
32. Tang, J. and Moret, B.M.E. 2003. Phylogenetic reconstruction from gene rearrangement data with unequal gene contents. In WADS ‘03. LNCS 2748, Springer Verlag, pp. 37–46.
33. Tesler, G. 2002 GRIMM: genome rearrangements web server. Bioinformatics, 18, 492–3.
34. Tesler, G. 2002. Efficient algorithms for multichromosomal genome rearrangements. Journal of Computer and System Sciences 65:587–609.
35. UCSC Genome Browser
36. Ware, D., Jaiswal, P., Ni, J., Pan, X., Chang, K., Clark, K., Teytelman, L., Schmidt, S., Zhao, W., Cartinhour, S., McCouch, S. and Stein, L. 2002. Gramene: a resource for comparative grass genomics. Nucleic Acids Research 30, 103–5.
37. Zheng, C., Lenert, A. and Sankoff, D. 2005. Reversal distance for partially ordered genomes. Bioinformatics 21, in press
38. Zheng, C. and Sankoff, D. 2005. Genome rearrangements with partially ordered Chromosomes. In Wang, L. (ed), Proceedings of COCOON ‘05. LNCS 3595. Berlin, Heidelberg:Springer Verlag, in press.