Original Synteny

Vincent Ferretti[1] *, Joseph H. Nadeau[2], David Sankoff[1] **

[1] Centre de recherches mathématiques, Université de Montréal,
CP 6128 Succursale Centre-ville, Montréal, Québec, H3C 3J7
[2] Department of Human Genetics, Montreal General Hospital,
McGill University, Montréal, Québec, H3G 1A4
and Jackson Laboratory, Bar Harbor, Maine 04609

Abstract. The inference of genome rearrangement requires detailed gene maps of related species. For most multichromosomal species, however, knowledge of chromosomal assignment of genes outstrips mapping data. Comparison of these species is thus a question of comparing sets of syntenic genes, without any gene order or gene orientation information. Given synteny data from present-day species, can we infer the synteny sets of ancestor species? How many chromosomes did these ancestors possess, and what genes were on each one? We first study the problem of calculating a syntenic edit distance between two genomes, based on reciprocal translocation, chromosome fusion and fission. This distance is then used in the analysis of the median problem for synteny, and hence for a prelimary approach to phylogenetic inference of synteny.

1 Introduction

The recent interest in the inference of genome rearrangement is inspired by the ongoing construction of detailed gene maps of a few model organisms and organelles [4, 2, 1]. For many more multichromosomal species, however, knowledge of chromosomal assignment of genes far outstrips any data on precisely where they are located on the chromosomal map. Two genes are said to be *syntenic* in an organism if they are assigned to the same chromosome. Comparison of species without maps is thus a question of comparing sets of syntenic genes, without any gene order or gene orientation information [5].

Since such intrachromosomal events as inversion and local transposition do not affect chromosomal assignment, they are not inferable from synteny data alone. Only interchromosomal events such as reciprocal translocation, chromosome fusion and chromosome fission affect synteny and are hence inferable from synteny data.

Given synteny data from present-day organisms, to what extent can we infer the synteny sets of ancestor organisms? How many chromosomes did these

[*] ferretti@ere.umontreal.ca

[**] Research supported by grants from the Natural Sciences and Engineering Research Council of Canada and the Canadian Genome Analysis and Technology program. Fellow of the Canadian Institute for Advanced Research; sankoff@ere.umontreal.ca

ancestors possess, and what genes were on each one? To answer these questions, we develop algorithms for a series of increasingly difficult problems. These algorithms are all computationally feasible for realistic data sets, all provide upper bounds on the true solutions, and all come demonstrably close to the true solutions as assessed by simulation and other means.

The first problem is that of calculating a syntenic edit distance between two genomes, based on translocation, fusion and fission.

This distance is then used in the analysis of the median problem for synteny, namely the construction of a genome the sum of whose distances to three given genomes is miminal.

The last problem is that of the optimization of the internal vertices of a given phylogenetic tree, given synteny data from species representing each of the terminal vertices. Capitalizing on the method developed for this problem, we also sketch an approach to phylogenetic inference based on synteny data.

2 Reciprocal Translocation and Synteny Sets

For our purposes, a genome is a set of objects called genes, partitioned into k subsets called synteny sets or chromosomes. A reciprocal translocation is an operation that transforms two chromosomes A and B into $(A - A') \cup B'$ and $(B - B') \cup A'$, respectively, where at least one of A' and B' is a proper subset of A or B. A fusion occurs when, e.g. $A' = A$ and $B' =$ the null set, and a fission when either A or B is replaced by the null set, in this formulation.

The question becomes one of calculating how many operations - translocations, fusions and fissions - it takes to convert one given genome to another.

3 An Upper Bound Algorithm for Syntenic Distance

Relabeling each gene in Genome 2 by its homologue's chromosome assignment in Genome 1, and collapsing duplicate instances of a label in a chromosome, we achieve a compact represention of the problem, e.g.:

```
Genome 1:
Chromosome 1:{x,y} Chromosome 2:{p,q,r} Chromosome 3:{a,b,c}

Genome 2:
Chromosome 1:{p,q,x} Chromosome 2: {a,b,r,y,z}

Compact representation of problem: {1,2},{1,2,3}
```

Note that genes, such as c and z above, whose chromosomal assignments are known in only one of the genomes, are ignored in the compact representation and are not taken account of in the distance calculation. In contrast to some other sequence comparison problems, the absence of such genes does not represent a deletion or insertion process, simply a lack of scientific information about which chromosome contains the gene, since almost all genes in one mammal, say, will have a counterpart in any other mammal.

A solution consists of a series of translocations, fusions and fissions that transform Genome 2 to the k chromosomes of Genome 1, i.e. to $\{1\}, \{2\}, ..., \{k\}$ (since we can assume without loss of generality that all the genes in the one are also present in the other). E.g.:

```
{1,2},{1,2,3} transformed by translocation to {1},{2,3}
{1},{2,3} transformed by fission to {1},{2},{3}

distance = 2
```

The other algorithms developed here loop extensively on the distance calculation so this has to be extremely time-efficient. The following heuristic is rapid and, as will be demonstrated below, is relatively accurate.

At each step, a first calculation finds a single label l to focus on, and the second determines the transformation it is involved in. We will describe them in reverse order.

Suppose l appears in $r(l)$ chromosomes of Genome 2.

If $r(l) = 1$ and some of labels syntenic with l appear in no other chromosome, we simply effect a fission to create a separate chromosome $\{l\}$.

If $r(l) = 1$, and all of the labels l' syntenic with l appear in $r(l') \geq r_{\text{min}} > 1$ chromosomes, we effect a translocation, again to obtain a separate $\{l\}$ chromosome. Here, the second chromosome involved in the translocation contains a label l', syntenic with l, with $r(l') = r_{\text{min}}$, and, if there are several such, with a maximal number of terms syntenic with l. E.g.:

```
{1,2,3,4},{2,3,5},{2,3,4},{4,5,6},{4,8,9}
```

if $l = 1$, $r_{\text{min}} = 3$, $l' = 2$ or $l' = 3$, and the second chromosome in the translocation is $\{2, 3, 4\}$, producing:

```
{1},{2,3,4},{2,3,5},{4,5,6},{4,8,9}
```

If $r(l) > 1$, we effect $r(l) - 2$ fusions followed by one translocation, again to produce a separate $\{l\}$.

Knowing how to act on any specific label, we return to the original question, which label l should we focus on first? We establish the following priorities:

1) any l for which $r(l) = 1$

2) any l for which $r(l) = 2$

3) if all $r(l) > 2$, pick l which minimizes $r(l)$ and, if there are several such, which minimizes $r(l')$ for some label l' in the chromosome remaining from the last operation involving l. If there are several such, choose l so that after it is operated on, $\sum_l r(l)$ is minimized.

The above procedure, while we have no guarantee of its optimality, is designed to choose an operation at each step in such a way as to set up the most advantageous situation possible for the next step.

3.1 Simulations and Tests

The first test we designed is a validation, albeit indirect, of the accuracy of the method. To the extent that the algorithm produces the true minimum, then the result of converting Genome 1 to Genome 2 should be the same as the reverse operation, since the inverse of every translocation is a translocation, the inverse of every fission is a fusion and the inverse of every fusion is a fission. If the algorithm is a poor approximation, we might expect the results in the two directions to differ frequently. Based on 20,000 randomly constructed genome pairs designed to model data on mammals (17-26 chromosomes containing 17-26 genes each), 65 % of the time the results of applying the algorithm in the two opposing directions were identical, 34 % of the time they differed by exactly 1, and in only 1 % of the cases did they differ by 2 (or more). Of course, in some of the zero-difference cases both directions might have given identical suboptimal results, but this is unlikely to have exceeded 3 % of the cases, and the results are extremely unlikely to have been off by more than 1. In all further uses of the distance algorithm, it was applied in both directions and the minimum value chosen.

To test the relation of syntenic distance to evolutionary history, we generated random genomes by applying series of random translocations to $\{1\}, ..., \{k\}$. For numbers of translocations less than about $\frac{k}{2}$, the algorithm reconstructed the right number of translocations, and never too many. As the number of translocations increased, the algorithm tended to underestimate the true number, finding shorter distances between the random genomes and $\{1\}, ..., \{k\}$.

These results are summarized in Table 1.

3.2 Complexity Results

Das Gupta, Jiang, Kannan, Li and Sweedyk [6] have proved the following results on the syntenic distance:

- The synteny problem is NP-complete.
- There is an approximation algorithm that achieves a factor of 2 approximation.
- There is an algorithm that runs in time $O(d^d k^c)$ to compute the syntenic distance exactly when this distance is d. (Useful when d is small.)
- There is an ordering of the given sets such that the following algorithm performs at most $d + \log d$ moves:

Let the input sets be $S_1, S_2, \ldots, S_k$ in the chosen ordering.

```
$Currentset = \{\}$
For $i = 1$ to $k$
    if there is an $x$ in $Currentset$ or in $S_i$ and $x$ does not occur in $S_j$ for $j > i$
            Perform the translocation on $Currentset$ and $S_i$ yielding:
            the singleton set $\{x\}$
            and $Currentset = \{Currentset \cup S_i\} - \{x\}$.
    else
            $Currentset = Currentset \cup S_i$.
```

Table 1. Underestimation of number of translocations generating a genome

number of chromosomes	number of trans-locations	mean distance calculated
10	1	1
	4	3.86
	7	6.53
	10	7.4
	13	8.6
	16	9.53
	19	10.33
15	1	1
	4	3.87
	7	6.33
	10	9
	13	10.8
	16	12
	19	13.33
	22	13.6
20	1	1
	4	3.73
	7	6.6
	10	9.13
	13	11.3
	16	13.6
	19	15.4
	22	16
	25	17.53
25	1	1
	4	4
	7	6.53
	10	9.13
	13	11.47
	16	14.47
	19	15.6
	22	17.8
	25	19.2
	28	21
	31	22.47

- For any constant factor approximation of the bipartite clique problem, a $(1+\epsilon)$ approximation can be obtained for the syntenic distance.

4 The Median Problem

Let $d(\text{Genome 1}, \text{Genome 2})$ be the syntenic distance between Genome 1 and Genome 2. Consider the simplest problem of phylogenetic inference, namely, the inference of the ancestral state in an unrooted tree with three terminal vertices. This is the *median* problem, defined for synteny data as follows: Given three genomes $1, 2$ and 3, we are required to construct a genome S such that

$$d(S,1) + d(S,2) + d(S,3).$$

is minimized. For this to be non-trivial in the context of our syntenic distance, Genome S must be constrained to contain certain genes, at least those that are identified in all of Genomes 1,2 and 3, or in two out of the three genomes, or possibly in any of the three genomes. Otherwise, if S can be empty, then the sum of the three distances is trivially zero. However this constraint is formulated in any particular context, we refer to it as the Median Content Constraint (MCC). In Section 5 below, we discuss how this constraint is the key to phylogenetic analysis based on synteny data.

4.1 A Simple Heuristic

As a preprocessing step, for any set of genes that are syntenic in all three data species, it suffices to consider only one exemplar in the analysis. (The full set is restored after the algorithm has terminated.) Similarly for any set of genes that are syntenic in two species and unidentified in the other.

As a first step, we choose any gene which must be in S according to the MCC, and define the initial chromosome in S to be one containing this gene.

Then while there remain unassigned genes satisfying the MCC, if there is one that can be added to an existing chromosome in S or which can initiate a new chromosome, without adding to the current overall cost, we do so. If not, we assign a gene to a pre-existing ancestral chromosome, or to a new chromosome, according to which gene, and which assignment, minimizes the sum of the distances to the terminal nodes based on the partial gene set assigned to date. This may not be unique, and we may thus be obliged to work on several alternative solutions in parallel. I.e., there may be several such minimizing genes and minimizing chromosomal assignments, but we retain for further exploration only those solutions involving one such gene having a minimum number of alternative assigments. We also have an option of limiting the number of solutions being constructed in parallel (in practice, 10).

The final assignment can then be improved by trying to move each gene in turn to a different chromosome and recalculating the sum of the three distances, iterating until no further improvement is possible.

The algorithm attains a local minimum depending on the first gene chosen and the starting points for other searches during execution, as well as the optional limit on the number of stored partial solutions. The algorithm is repeated a number of times (in practice, 20) in the hope that the best local minimum will be a global minimum.

5 Optimizing a Given Phylogeny

A method for the inference of ancestral synteny can be adapted from the iterative improvement method of [3]. We assume that the given phylogeny is an unrooted binary tree on n terminal nodes, each of which is associated with some real genome. Consider the $n-2$ 3-stars defined by each internal node and its three neighbors. A most parsimonious solution will have a reconstructed genome associated with each internal node, and each one must be a solution to the median problem determined by its neighbors.

We can try to find such a solution by starting with some initial solution and iteratively improving the stars on three vertices, by the heuristic for the median. This process will eventually converge to a local optimum.

5.1 Tests on Mammalian Genomes

We extracted synteny data from the query system of the Mouse Genome Databank, for eleven mammalian genomes, where the number of genes identified as to chromosomal assignment ranges from a few dozen (cat, sheep, Chinese hamster), to a few thousand (human, mouse).

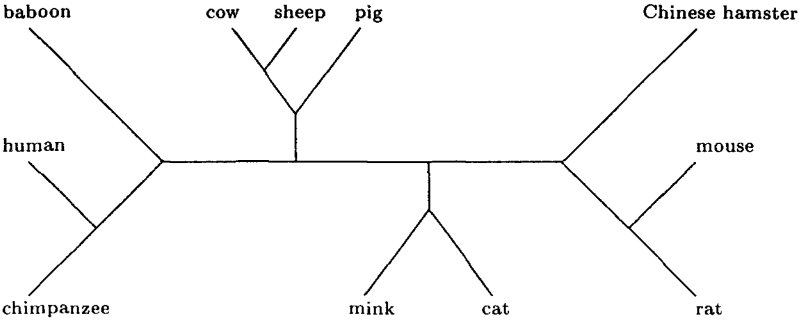

Fig. 1. Phylogenetic tree assumed to have generated comparative synteny data on mammalian genomes

We assumed the phylogeny in Figure 1 and undertook to reconstruct the ancestral genome syntenies through the above method. Had all these species many identified genes in common, it would have sufficed to impose a weak MCC where only genes occurring in all three of the genomes were considered in the various median calculations. For the time being, however, there are not enough genes homologous in enough species throughout the mammalian phylogeny to use this MCC; some internal vertices could not be reconstructed at all. The incorporation of genes with homologues in only two of the three neighbors solves this problem. Indeed we attempted to use the strongest MCC possible; to include all those genes occurring in only one of the three genomes as well, but this led to an intolerable explosion in the number of alternative solutions to be considered. Thus we introduced a lightly attenuated constraint: include those genes in only one of the three genomes if they can be added after all the other genes are assigned chromosomes, *in only one cost-free way.* Thus for genes identified only in human and mouse, say, this MCC conceivably allows their syntenies to interact at some internal vertices, where the weaker MCCs would necessarily simply discard the evidence of these genes.

As an initial solution, we assigned to each internal vertex a genome copied from a close terminal vertex. Each iteration of the algorithm thus involves the solution of 9 median problems. Table 2 summarizes the progress of the algorithm from iteration to iteration as it converged to a final solution of 109 translocations, fusions and fissions.

Table 2. Progress of algorithm in converging to local optimum in terms of number of chromosomes in ancestral genomes and total syntenic distance. Note that in first iteration, genes occurring only once in the neighborhood of an internal vertex have no effect on total cost, while in successive iterations many of these genes now occur twice in the neighborhood of the same internal vertices and do contribute to the cost. Note also that our relatively strong MCC results in hundreds of genes in each ancestral genome, though only a few dozen may be known in the species associated with neighboring terminal vertices

ancestral	iteration					number of genes
nodes	1	2	3	4	5	in genome
human-chimp	22	23	23	23	23	774
primates	22	25	24	24	24	666
sheep-cow	33	31	31	31	31	403
artiodactyls	25	32	31	31	31	568
primates-artiodactyls	21	25	31	31	31	693
carnivores	17	17	19	19	19	606
carnivores-rodent	18	20	20	20	20	635
rodent	19	19	19	19	19	482
mouse-rat	19	19	19	19	19	528
total syntenic distance	103	112	110	109	109	

6 Phylogenetic Reconstruction

The method for optimizing a given phylogenetic tree is based on the minimum number of translocations, fusions and fissions necessary to account for the data, assuming that tree is correct. It is a conceptually simple step to proceed to find the most parsimonious tree by exhaustive evaluation of the set of possible trees for these data. Of course this is computationally not feasible for even a moderate size data set because of the exponential growth of the number of different trees as a function of the number of terminal vertices, i.e. the number of species.

Our approach is feasible, however, for evaluating a limited number of competing hypotheses. As an illustration of this principle, we applied the method to the data treated in Section 5, this time assuming a tree derived from that in Figure 1 by rotating the chimpanzee into the position occupied by the rat, the rat into the pig's position, and the pig into the chimpanzee's position. This tree required 117 translocations, fusions and fissions, clearly not as parsimonious as the tree in Figure 1.

7 Further Work

The problems addressed in this paper should admit better solutions than we have been able to find in this preliminary investigation, e.g., an exact algorithm for the syntenic distance problem for realistic values of k, namely $k \approx 25$. The median problem is probably much harder, but improvement permitting efficient calculation under the strong MCC should be possible.

References

1. S. Hannenhalli and P.A. Pevzner. Transforming cabbage into turnip. (polynomial algorithm for sorting signed permutations by reversals). *Proceedings of the 36th Annual Symposium on Foundations of Computer Science,* 1995.
2. J. Kececioglu and D. Sankoff. Exact and approximation algorithms for sorting by reversals, with application to genome rearrangement. *Algorithmica,* 13:180-210, 1995.
3. D. Sankoff, R.J. Cedergren, and G. Lapalme. Frequency of insertion-deletion, transversion, and transition in the evolution of 5S ribosomal RNA. *Journal of Molecular Evolution,* 7:133–149, 1976.
4. 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.
5. I. A. Zakharov. Measurements of similarity of synteny groups and an analysis of genome rearrangements in the evolution of mammals. In *Bioinformatics, supercomputing and complex genome analysis,* H. Lim, J. Fickett and C. Cantor (eds), World Scientific, 1993.
6. B. Das Gupta, T. Jiang, S. Kannan, M. Li, and Z. Sweedyk. Personal communication, March 1996.