Conserved segment identification

David Sankoff * Vincent Ferretti † Joseph H. Nadeau ‡

1 Introduction.

The study of genome rearrangements based on map data depends crucially on the definition and identification of conserved segments, regions of chromsomes in two related species in which both gene content and gene order are parallel in the two species, as illustrated in Figure 1(a). As map data accumulate, however, it becomes increasingly difficult to find segments that satisfy the criteria of content and order perfectly. This can be attributed, in unknown proportions, to experimental error - either gross mistakes in chromosomal assignment of genes or quantitative errors in map positions affecting apparent gene order - or to relatively high rates of inversion and transpositions of small regions of chromosomes. In the human-mouse comparison, stringent requirements of parallel content and order lead to a proliferation of short segments inferred instead of the few long segments which have traditionally been recognized. This is illustrated in Figure 1(b).

To correct this problem, our aim will be to try to recover, insofar as possible, the configuration of conserved segments that results from the evolutionary history of reciprocal translocations accounting for the gross differences between the genomes. Our hypothesis is that this goal can be achieved with some accuracy by minimizing appropriately weighted mapping error plus rearrangement costs. As formulated, this is carried out using a variant of single link stepwise cluster analysis performed simultaneously on all conserved synteny sets (sets of genes occurring in common on one human chromosome and one mouse chromosome), with the interim results from each cluster analysis affecting the current state of all other cluster analyses. Parameters are varied so that the solution approaches in general characteristics, if not in detail, consensus reconstructions arrived at by experts.

*Centre de recherches mathématiques, Université de Montréal, CP 6128 Succursale Centre-ville, Montréal, Québec H3C 3J7. E-mail: sankoff@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.

†Centre de recherches mathématiques, Université de Montréal, CP 6128 Succursale Centre-ville, Montréal, Québec H3C 3J7. E-mail: ferretti@ere.umontreal.ca.

‡Jackson Laboratory, Bar Harbor, Maine 04609. Present address: Department of Genetics, Case Western Reserve University, Cleveland, Ohio 44106. E-mail: jhn4@po.cwru.edu.

RECOMB 97, Santa Fe New Mexico USA

2 Background.

The quantitative approach to partitioning two genomes into corresponding pairs of conserved segments was initiated in [4]. This paper made explicit the hypothesis that the observed configuration of conserved segments is essentially due to repeated, random occurrences of the process of reciprocal translocation. Recent updates of this approach are found in [1, 2]. Mathematical extensions of the random translocation model can be found in [3, 5, 6].

The extent of the problem of inaccurate map position can be seen in [2]. As for experimental error in the assignments of genes to chromosomes, some of this is due to incorrect homology decisions involving sets of duplicate genes. The following statistics are revealing. In April 1996, the Mammalian Genome Database contained 28 genes which each constituted the sole evidence of a homologous segment in some human chromosome and some mouse chromosome, out of about 110 conserved syntenies in all. By August 1996, five of these genes had been removed from either the human or mouse data, four had been reassigned in one or both of the genomes, and only two had been confirmed by the report of another gene on both the human and mouse chromosomes. An additional 6 single-gene segments also appeared in the database at this date.

3 The objective function.

The smallest number of segments - subgroupings of conserved syntenic genes - that can be produced by any analysis is just the total number of conserved synteny sets $c \leq c_1 c_2$, where c_1 and c_2 are the number of chromosomes in species 1 and species 2, respectively. This solution is generally not acceptable because it groups all genes belonging to a conserved synteny, no matter how dispersed they are along the chromosome, into a single conserved segment, and it does not allow for the real possibility that a single conserved synteny may be the result of two or more translocation events. At the other extreme, the largest number of segments that can possibly be obtained is n, the total number of homologous genes identified in the two genomes, simply by assuming that each gene defines a different conserved segment and that genes are adjacent in two genomes only by coinci-

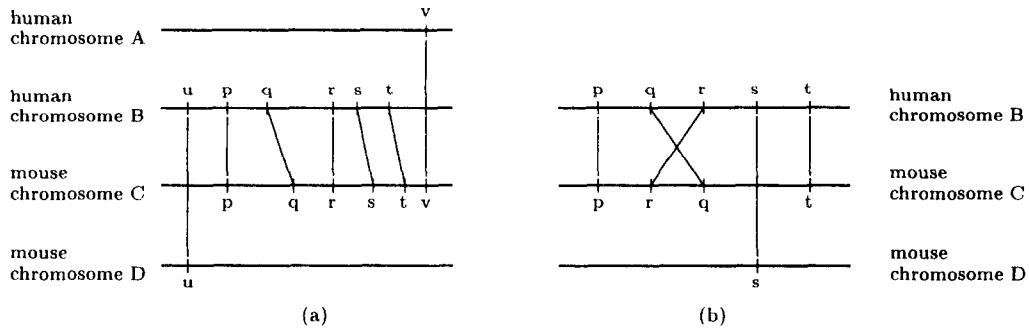

Figure 1: (a). Schematic example of conserved segment in a human chromosome (B) and a mouse chromosome (C). Genes u and v have homologues elsewhere in the mouse and human genomes, respectively, and thus limit the leftward and rightward extension of the segment.
(b). Experimental mistake in the chromosomal assignment of s to mouse chromosome D, quantitative error in the assignment of q and/or r in the human or mouse map, or inversion of qr or transposition of q or r, results in the erroneous identification of three segments, p,qr,t, instead of just one, in human chromosome B and mouse chromosome C, and an additional one, s, in human chromosome B and mouse chromosome D.

dence. This solution is even less realistic than the opposite extreme. More interesting solutions lie somewhere between these two extremes. For an appropriate choice of weighting parameters, α, β, γ, we wish to find the subgroupings of conserved syntenic genes into b segments, for $c \leq b \leq n$, so as to minimize

$$D = \sum_{i=1}^{b} D_i,$$

where D_i is a weighted measure of the compactness, density and integrity of segment i. Compactness is determined by how close together, in metric terms, the genes in a segment are located in both species. Operationally, this concept will be realized by the maximum distance between any two genes in the human segment plus the maximum distance between any two genes in the mouse segment. Density can be assessed simply by counting how many genes are in a segment and comparing this to its metric length. Integrity of segment i is measured by how many other segments have elements intervening, in one or both species, between members of i. Formally,

$$D_i = \gamma \max_{x,y \epsilon i(1)} |x - y| + \alpha s[i(1)]$$

$$+\gamma \max_{x,y \epsilon i(2)} |x - y| + \alpha s[i(2)] - \beta m(i),$$

where $x \epsilon i(j)$ refers to a gene (or its map coordinate) in segment i in species j, $m(i)$ indicates the number of homologous gene pairs in segment i and $s[i(j)]$ denotes the number of *other* segments with elements within the range of segment i in species j.

Note that, as formulated, the density term is superfluous, since $\sum m(i) = n$, a constant. In Section 4, however, we will see that the inclusion of this term in a stepwise algorithm privileges certain intuitively more plausible solutions over others with the same value of D.

4 The algorithm.

Direct minimization of $D = \sum D_i$ is generally not feasible, because what is included in segment i impacts the quality of other segments and vice-versa. Instead we propose a rapid stepwise upper-bound algorithm and show sufficient conditions for it to calculate D exactly. An advantage of this method is that it constructs solutions for all b in one pass.

Our procedure starts with the extreme solution where $b = n$, the total number of homologous genes in the analysis. We than combine step by step genes syntenic in both genomes into conserved segments, starting with those genes that are closest together in both genomes. Because integrity depends on the number s of other segments intervening in a given segment, and not the number of genes in these other segments, each step in the analysis of the $i-$th set of conserved syntenic genes, by decreasing the number of segments by 1, may affect, through the s terms in D_j, the further analysis of the j-th conserved synteny.

Basic to the algorithm is the notion of a rooted binary branching tree T_i with the leaves, or terminal nodes, associated with the n_i genes in conserved synteny i. This is illustrated in Figure 2. Each nonterminal node v denotes the formation of a segment from two smaller segments v_1, v_2 of distance $d(v) = d(v_1, v_2)$ apart, where

$$d(v_1, v_2) = \gamma[|x_1 - x_2| + |y_1 - y_2|]$$

$$+\alpha[s(x_1, x_2) + s(y_1, y_2)]$$

$$-\beta[m(v_1) + m(v_2)],$$

where $x_i \epsilon v_i$ and $y_i \epsilon v_i$ refer to genes in segment v_i at positions x_i and y_i, in species 1 and species 2, respectively, which minimize $|x_1 - x_2|$ and $|y_1 - y_2|$. $m(v_i)$ is the number of genes in segment i and $s(z_1, z_2)$ denotes the number of different segments with elements between the genes at z_1 and z_2. Note that d is a more general type of distance score than a metric, and it is defined only for two segments v_1 and v_2 containing genes in the same synteny sets.

The data on the map location of genes generally has an implicit and small range of uncertainty or an explicit and larger range. The distance $d(x_1, x_2)$ is defined to be the minimum possible given the ranges of the two genes x_1 and x_2. After precalculating all these distances, we apply the following:

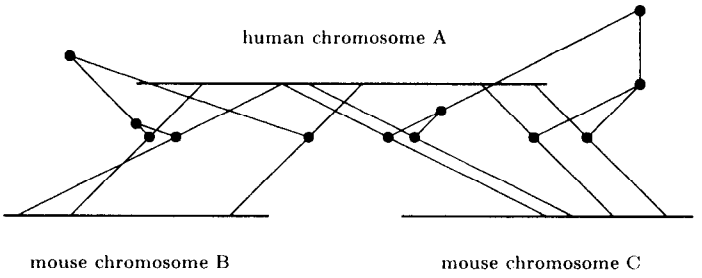

Figure 2: Two rooted binary trees each representing successive solutions to the problem of identifying conserved segments within two conserved syntenies. Thin lines connect homologous genes in the two genomes. Note that the conserved syntenies overlap on the human chromosome and that the number of segments from the synteny on the right intervening between genes on the left changes as the trees are constructed from bottom up.

Algorithm conseg

Let n_k be the number of genes in the k-th conserved synteny. Set $b = n = \sum n_k$, the total number of homologous pairs of genes, and let *seg* to be the set of all these genes. For all k, set $S_k = -\beta n_k$. Initial construction step for T_k: Identify the terminal nodes with the n_k genes in the conserved synteny.

While there remains a conserved synteny made up of two or more segments in *seg*,

> Find the two segments v_1 and v_2 that are closest together, i.e. that minimize $d(v_1, v_2)$.
>
> Combine v_1 and v_2 to form a new segment v. Add v to *seg* and remove v_1 and v_2 from *seg*. If either or each of v_1 or v_2 is a single gene, fix its position (on both chromosomes) within its range to be consistent with $d(v_1, v_2)$.
>
> If v contains genes in the k-th synteny, update the construction of T_k to indicate the branching of v to v_1, v_2 and set $S_k = S_k + d(v) + \beta[m(v_1) + m(v_2)]$.
>
> Set $b = b - 1$, and output configuration of the b segments in *seg*.
>
> Recalculate all distances d given the decrease in number of segments in *seg* and the possibly newly fixed position of one or two genes.

Endwhile
Set $D^* = \sum S_k$.

As presented, this algorithm is greatly simplified. For example, when the segments chosen to combine overlap, it is sometimes necessary to forgo fixing the positions of the genes within them until a later time. And special measures must be taken when many genes are mapped to the same point. But these situations may be incorporated without changing the basic concepts of the algorithm.

The role of the density parameter β becomes clear in this algorithm. Rather than combining single genes or relatively sparse segments of a certain length, there is a bias towards combining, whenever possible, relatively dense segments of the same length. This ensures that the most clearcut examples of conserved segments emerge early during the execution of the algorithm and are present for as wide a range of b as posssible.

The clustering procedure may seem a roundabout way of approaching the objective function, but to the extent that gene order is conserved within segments, the following theorem becomes pertinent.

Theorem The upper bound D^* achieved by the algorithm is equal to the objective D if for each node $v = v_1 \cup v_2$, the genes in v_1 and the genes in v_2 are in disjoint intervals in both genomes.

5 Application

Initial applications of the **conseg** algorithm, which has been implemented in C, to human/mouse homologies gives results comparable to the published works of experts, e.g. in [1, 2]. Figure 3 illustrates the results for mouse chromosome 4 for two values of b, compared to the analyses in the two sources.

Table 1 shows how the output of **conseg**, with parameters suitably adjusted, can conform relatively well, in terms of number of segments per chromosome, to the judgements of experts using quite different standards for identifying segments. Though some discrepancies (e.g. in mouse chromosomes 3,7,19) are no doubt due to special biological circumstances not taken into account by **conseg**, others are likely due to expert's variable application of subjective criteria from chromosome to chromosome. In addition, the segments presented in [1] do not take full account of segment disruption due to intervening segments within the human chromosome. Finally, our data set is more recent than that in [2], which in turn is more recent that in [1].

6 Discussion

The results of the analysis for a fixed value of b represent, *grosso modo*, a hypothesis about the rearrangement events resulting in the current configurations of conserved segments. A segment X which is interrupted by other segments is presumed to have incurred these interruptions through intrachromosomal events, either before or after the translocation which gave rise to X. Segments Y and Z which are analyzed as distinct although they are in the same conserved synteny are presumed to have arisen through separate translocation events.

Thus an analysis resulting in a higher value of b implicitly assumes more interchromosomal exchanges, i.e. con-

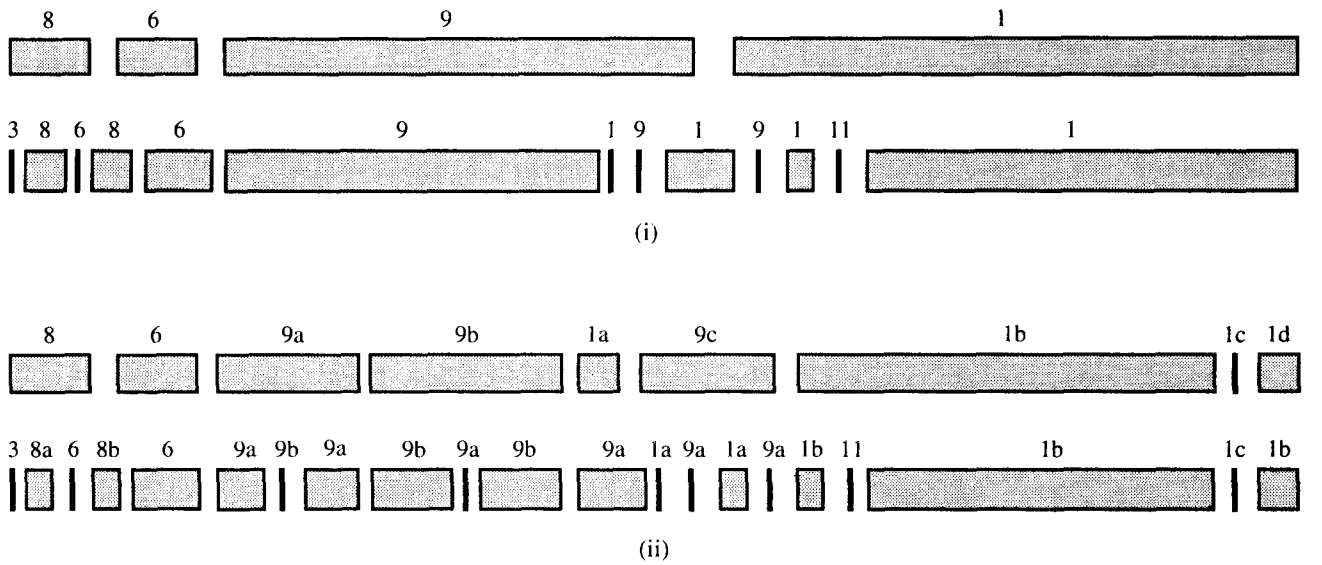

Figure 3: Mouse chromosome 4: Human chromosome numbers corresponding to segments are indicated. In each analysis, only lengths of chromosome with the same human chromosome number *and* letter code are considered to belong to the same segment.
(i). Analysis in [1] compared to **conseg** analysis for $b = 113$, where $\alpha = \gamma = 1$ and $\beta = 0.3$.
(ii). Analysis in [2] compared to **conseg** analysis for $b = 183$, where $\alpha = 15, \gamma = 1$ and $\beta = 0$.
Note that the **conseg** solutions involve more discontinuous segments than those of [1] and [2]. In [1], much fewer data were available and compact segments on the mouse chromosomes were generally retained in spite of the non-adjacencies of their human homologs. In [2], the map positions of many markers were adjusted to produce non-overlapping segments. Both analyses ignored genes whose chromosome assigments seemed dubious.

Mouse chromosome	Copeland et al. [1]	b=113 $\alpha = 1$ $\gamma = 1$ $\beta = 0.3$	Debry, Seldin [2]	b=183 $\alpha = 15$ $\gamma = 1$ $\beta = 0$
1	5	7	11	12
2	9	8	9	11
3	8	5	8	13
4	4	6	9	10
5	5	7	11	13
6	5	6	11	13
7	9	6	20	11
8	11	7	9	9
9	7	6	11	10
10	6	6	13	10
11	6	8	10	9
12	3	3	5	5
13	5	6	9	11
14	8	5	9	8
15	4	3	4	4
16	4	5	9	7
17	8	8	10	11
18	4	4	9	8
19	3	6	4	8
Total	115	113	183	183

Table 1: Number of conserved segments in two analyses of human/mouse data, compared to **conseg** output.

served syntenies containing several segments arise from multiple translocations. Analyses characterized by lower values of b attribute the disruption of conserved syntenies by intervening segments to intrachromosomal events. Statisical analyses of the number of syntenic segments versus the total number of conserved segments on a chromosome, in comparison with a random translocation model, should help delimit a reasonable range of values for b.

Another aspect, which we have only begun to explore, concerns appropriate values for the weighting parameters. There are clear limits e.g. β/γ should be less than overall segment density. The approach we presently follow is to compare the number U_i of different human chromosomes represented among the b_i segments on a single mouse chromosome i, with the number u_i expected under a random hypothesis:

$$u_i = 22[1 - (\frac{21}{22})^{b_i}].$$

We chose the parameter values and b so that

$$\sum_{i=1}^{19} u_i = \sum_{i=1}^{19} U_i.$$

In our data set, these values are $b = 174$, $\alpha = 10$, $\gamma = 1$ and $\beta = 0.3$.

7 Acknowledgements

We thank Marge May of Jackson Labs for help in extracting conserved synteny data from the Mammalian Genome

Database. We also acknowledge helpful comments received from several participants at the The University of Pennsylvania Conference on Computational Biology to honor the 50th anniversary of the ENIAC, held at Princeton University in May 1996, where an earlier version of this work was presented.

References

[1] N.G. Copeland, N.A. Jenkins, D.J. Gilbert, J.T. Eppig, L.J. Maltais, J.C. Miller, W.F. Dietrich, A. Weaver, S.E. Lincoln, R.G. Steen, L.D. Stein, J.H. Nadeau and E.S. Lander. A genetic linkage map of the mouse: current applications and future prospects. *Science*, 262: 57-66, 1993.

[2] R.W. Debry and M.F. Seldin Human/mouse homology relationships. *NCBI Web Site*, http://www3.ncbi.nlm.nih.gov/Homology/.

[3] V. Ferretti, J.H. Nadeau and D. Sankoff. Original synteny. *Proceeding of the Seventh Annual Symposium on Combinatorial Pattern Matching*, D. Hirschberg and G. Myers ed., Springer Verlag Lecture Notes in Computer Science, 1075: 159-167, 1996.

[4] J.H. Nadeau and B.A. Taylor Lengths of chromosomal segments conserved since divergence of man and mouse. *Proceedings of the National Academy of Sciences USA*, 81: 814-818, 1984.

[5] D. Sankoff and V. Ferretti. Karyotype distributions in a stochastic model of reciprocal translocation. *Genome Research*, 6: 1-9, 1996.

[6] D. Sankoff and J.H. Nadeau. Conserved synteny as a measure of genomic distance. *Discrete Applied Mathematics*, (in press).