Preprint
For final version ©2003 Cold Spring Harbor Laboratory Press, see
Genome Research **13** (2003), 37–45.

Genome Rearrangements in Mammalian Evolution: Lessons from Human and Mouse Genomes

Pavel Pevzner and Glenn Tesler

Department of Computer Science and Engineering,
University of California, San Diego, CA

Abstract

Although analysis of genome rearrangements was pioneered by Dobzhansky and Sturtevant 65 years ago, we still know very little about the rearrangement events that produced the existing varieties of genomic architectures. The genomic sequences of human and mouse provide evidence for a larger number of rearrangements than previously thought and shed some light on previously unknown features of mammalian evolution. In particular, they reveal that a large number of micro-rearrangements is required to explain the differences in draft human and mouse sequences. Below we describe a new algorithm for constructing synteny blocks, study arrangements of synteny blocks in human and mouse, derive a most parsimonious human-mouse rearrangement scenario, and provide evidence that intrachromosomal rearrangements are more frequent than interchromosomal. Our analysis is based on the human-mouse breakpoint graph, which reveals related breakpoints and allows one to find a most parsimonious scenario. Since these graphs provide important insights into rearrangement scenarios, we introduce a new visualization tool that allows one to view breakpoint graphs superimposed with genomic dot-plots.

1 Introduction

Analysis of genome rearrangements in molecular evolution was pioneered by Dobzhansky and Sturtevant, 1938 [7], who published a milestone paper with an evolutionary tree presenting a rearrangement scenario with 17 inversions for the species *D. pseudoobscura* and *D. miranda*. Every genome rearrangement study involves solving a combinatorial puzzle to find a series of genome rearrangements to transform one genome into another. Palmer and co-authors (Palmer and Herbon, 1988 [26]) pioneered studies of the shortest (most parsimonious) rearrangement scenarios and applied this approach to plant mtDNA and cpDNA. Since then, the analysis of the most parsimonious scenarios has become the dominant approach in genome rearrangement studies. For unichromosomal genomes, it usually amounts to analysis of inversions (also known as *reversals*), which are the most common rearrangement events. The problem of finding the minimum number of reversals to transform one unichromosomal genome into another is known as the *reversal distance* problem. For multichromosomal genomes, the most common rearrangements are reversals, translocations, fusions, and fissions, and the number of such rearrangements in a most parsimonious scenario is known as the *genomic distance* between multichromosomal genomes.

Finding the reversal distance is a difficult combinatorial problem. In the very first computational studies of genome rearrangements, Watterson et al., 1982 [38] and Nadeau and Taylor, 1984 [24] introduced the notion of a *breakpoint* (disruption of gene order) and noticed some correlations between the reversal distance

and the number of breakpoints (in fact, Sturtevant and Dobzhansky, 1936 [33] implicitly discussed these correlations 65 years ago!). The shortcoming of early genome rearrangement studies is that they considered breakpoints independently without revealing combinatorial dependencies between *related* breakpoints. The simplest example of related breakpoints are two breakpoints formed by a single reversal. Kececioglu and Sankoff, 1993 [13] were the first to recognize the importance of dependencies between breakpoints and to come up with an approximation algorithm for the reversal distance problem. The important result of Bafna and Pevzner, 1993 [1] is the construction of the *breakpoint graph*, which reveals related breakpoints and allows one to find the most parsimonious scenarios.

Based on the notion of the breakpoint graph, Hannenhalli and Pevzner, 1995 [10] developed a polynomial algorithm for the reversal distance problem, i.e., for computing a most parsimonious scenario to transform one unichromosomal genome into another. This approach was further extended to the genomic distance problem, i.e., finding a most parsimonious scenario for multichromosomal genomes under inversions, translocations, fusions, and fissions of chromosomes (Hannenhalli and Pevzner, 1995 [9], Tesler, 2002 [35]). However, these results, while useful, do not yet yield a meaningful estimate of the number of the rearrangement events on the evolutionary path from mouse to human. The problem is that the genomic sequences provide evidence for both micro-rearrangements (e.g., intrachromosomal rearrangements with a span below 1 Mb) and macro-rearrangements (e.g., intrachromosomal rearrangements of larger span as well as interchromosomal rearrangements). The existing rearrangement algorithms do not distinguish between these two types of rearrangements. Since some micro-rearrangements may be caused by fragment assembly errors, mixing micro- and macro-rearrangements within one rearrangement scenario may produce a distorted picture greatly influenced by the sequencing errors in draft genomic sequences. Another difficulty is an unreliable assignment of orthologs (false orthologs) that may create an impression of a rearrangement that never happened (Tatusov et al., [34]). The conserved gene order can also be disrupted by recent duplications and insertions (Hardison et al., 1997 [11]).

To address these complications we first describe a new approach to synteny block generation that separates micro- from macro-rearrangements. It allows one to study micro- and macro-rearrangements separately and to arrive at a new estimate of the number of macro-rearrangements that cover about 170 Myr of evolutionary distance between human and mouse. We also estimate the number of micro-rearrangements (but it remains to be seen to what extent this estimate is influenced by the fragment assembly errors) on the evolutionary path between mouse and human.

2 Synteny blocks

In a pioneering paper, Nadeau and Taylor, 1984 [24] introduced the notion of *conserved segments* (i.e., segments with preserved gene orders without disruption by rearrangements) and estimated that there are roughly 180 conserved segments in human and mouse. Later, Copeland et al., 1993 [22], DeBry and Seldin, 1996 [6], Waterston et al., 2002 [28] and Gregory et al. [32] confirmed these estimates. In the past decade, the progress in understanding the evolutionary history of entire genomes was mainly based on comparative genetic maps (O'Brien et al., 1999 [25]). However, these estimates suffer from low resolution of comparative maps in certain genomic areas. Current genomic sequences provide evidence that the human and mouse genomes are significantly more rearranged than previously thought. Moreover, they indicate that a large proportion of previously identified conserved segments are not really conserved since there is evidence of multiple micro-rearrangements in many of them (Mural et al., 2002 [21]). These micro-rearrangements were not visible in the comparative genetic maps that were used for defining ≈ 180 conserved segments in the past. We study *synteny blocks* instead of conserved segments. Intuitively, the synteny blocks are segments that can be converted into conserved segments by micro-rearrangements; see the GRIMM-Synteny algorithm below for a formal definition. The synteny blocks do not necessarily represent areas of continuous

similarity between two genomes. Instead, they usually consist of short regions of similarity that may be interrupted by non-similar regions and gaps. Most synteny blocks are subject to micro-rearrangements within these blocks.

We demonstrate that human and mouse genomes share 281 synteny blocks of size at least 1 Mb (shown in Fig. 1a) and that at least 245 rearrangements of these blocks occurred since the divergence of human and mouse. The positions of these blocks in the human and mouse genomes are given in supplementary materials. The largest synteny block in human is 79.6 Mb and the average block size 9.6 Mb. The largest synteny block in the mouse genome is 64.8 Mb in length and the average block size is 8.5 Mb.

The overall size of syntenic blocks is approximately 2,707 Mb in human and 2,397 Mb in mouse. The breakpoint regions (i.e., intervals between consecutive syntenic blocks) vary and may be as large as 23.2 Mb in human and 6.7 Mb in mouse. The average size of breakpoint regions is 668 Kb in human and 458 Kb in mouse. The overall size of the breakpoint regions equals 172 Mb in human and 119 MB in mouse (although some of these breakpoint regions may host shorter synteny blocks). There is evidence of at least 3170 micro-rearrangements (reversals) that happened within the synteny blocks (though many of them may be artifacts of incorrect assemblies). This very high estimate of the number of micro-rearrangements further confirms the conjecture that micro-rearrangements are more common than previously thought (Carver and Stubbs, 1997 [5], Puttagunta et al., 2000 [27], Thomas et al., 2000 [37], Kumar et al., 2001 [16]). In fact, this number does not even include the micro-rearrangements within synteny blocks shorter than 1 Mb.

3 From local alignments to synteny blocks

Given two genomic sequences, how one can construct synteny blocks? False ortholog assignments and micro-rearrangements make it non-trivial to find the analogs of synteny blocks (conserved gene clusters) even in shorter bacterial genomes (Fujibuchi et al., 2000 [8], Lathe et al., 2000 [18], Wolf et al., 2001 [40], Rogozin et al., 2002 [29]). In addition, human-mouse sequence similarities in non-coding regions (Koop et al., 1994 [15], Thomas et al., 2000 [37]) may further complicate ortholog assignments and make it difficult to apply the methods developed in bacterial genomics to construction of human-mouse synteny blocks.

Sankoff et al., 1997 [30] were the first to come up with an algorithm for synteny block generation. However, their approach was mainly intended for comparative mapping data. Below we describe a different approach that is geared toward genomic sequences. To construct the human-mouse synteny blocks, we start with bidirectional best local similarities (also called *anchors*) between human and mouse genomic sequences (Tatusov et al., 1997 [34], Mural et al., 2002 [21]). A number of software tools have recently become available to generate such anchors for entire mammalian genomes (Mayor et al., 2000 [20], Schwartz et al., 2000 [31], Ma et al., 2002 [19], Kent, 2002 [14]). We assume that a set of non-overlapping anchors (local alignments between two genomes) is given and the goal is to construct the synteny blocks based on these anchors. We study the same versions of draft human and mouse sequences and the same set of anchors that were used in Waterston et al., 2002 [28]. This set of anchors was provided by Michael Kamal at the Whitehead Institute and was generated by PatternHunter (Ma et al., 2002 [19]). The set consists of 558,678 anchors with alignment length ranging from 40 to 9647 nucleotides (the mean is 340). We emphasize that these anchors do not necessarily represent similarities within human and mouse genes but may also represent similarities between non-coding regions. This is a departure from the previous "gene order comparison" approach of genome rearrangement studies. It allows us to bypass the difficult issues of gene annotation and ortholog identification, which are not necessary for genome rearrangement studies. This approach may miss similarities between some genes at evolutionary distances where protein similarity still exists but DNA similarity has faded away. However, this is not a serious concern for the rather similar human and mouse genomes. Moreover, our approach can be generalized to handle both DNA and protein similarities in a unified framework.

We assume that human and mouse chromosomes are concatenated to form a single coordinate system. An anchor that starts at position h in the human genome and at position m in the mouse genome is described by its starting point (h, m) in 2-D. We remark that in reality the anchors are not points (h, m) but diagonals in 2-D described by coordinates (h, m) of an alignment start and the length of the alignment. Such a coordinate system is shown in Fig. 1a, with chromosomes dividing the plane into rectangles. We define the distance between two points (h_1, m_1) and (h_2, m_2) from the same chromosome pair (the same rectangle) as the Manhattan distance $|h_2 - h_1| + |m_2 - m_1|$. The distance between points from different chromosome pairs is defined as infinity. The distance between two anchors is defined as the distance between their closest ends.

Although the number of anchors may be very large (hundreds of thousands) one still can apply fast genome rearrangement algorithms (Tesler, 2002 [35]) to find a most parsimonious scenario to transform the order of anchors in human into an order of anchors in mouse. However, this scenario will likely be unrealistic since many anchors may correspond to false orthologs. Therefore, a technique to filter out false orthologs (even at the expense of filtering some real orthologs) is needed. False orthologs will often look like isolated points (or "small clusters") in a genomic dot-plot, while synteny blocks will be formed from clusters consisting of a larger number of points. Fig. 2a presents the genomic dot-plots for anchors from the X chromosomes (a blow-up of the X-X rectangle from Fig. 1a). A brief look at Fig. 2a reveals 16 clusters (Fig. 2b). Fig. 2c presents *rectified clusters* that ignore the details of the internal anchor arrangements in the clusters and represent every cluster as a diagonal. These rectified clusters are further combined into diagonals that correspond to 11 synteny blocks (Fig. 2d). Although the synteny blocks in Fig. 2d differ in sizes, the sizes of synteny blocks are irrelevant for genome rearrangement algorithms. Fig. 2e is a symbolic representation of synteny blocks as units of the same size, used in the construction of the breakpoint graph.

The above description hides many important details and in many cases the choice of synteny blocks is less obvious. Below we describe the GRIMM-Synteny algorithm for synteny block generation from a collection of anchors. The algorithm uses the *gap threshold* G and *minimum cluster size* C as parameters and works as follows:

GRIMM-Synteny algorithm

(1) Form an *anchor graph* whose vertex set is the set of anchors.

(2) Connect vertices in the anchor graph by an edge if the distance between them is smaller than the gap size G.

(3) Determine the connected components of the anchor graph. Each connected component is called a *cluster*.

(4) Delete "small" clusters (shorter than the minimum cluster size C in length).

(5) Determine the cluster order and signs for each genome.

(6) Output the strips in the resulting cluster order as synteny blocks.

GRIMM-Synteny finds 319 clusters in the human genome that are longer than 1 Mb. In addition to these clusters we identified a number of smaller clusters; for example, in the human genome there are 36 clusters whose length varies from 0.5 Mb to 1 Mb, 21 clusters whose length varies from 250 Kb to 500 Kb, and 774 clusters with length from 50 Kb to 250 Kb. However, smaller syntenic blocks assignments are less reliable since they may be caused by false orthologs and sequencing errors.

Fig. 1b presents examples of some highly rearranged clusters from human chromosome 18/mouse chromosome 17 and the corresponding anchor graph. After constructing the cluster graph and deleting small

clusters (steps 1–4), one has to determine the cluster order and signs (step 5). We define the *span* of a cluster in human (mouse) as the interval between its minimum and maximum coordinates in human (mouse). Similarly to Mural et al., 2002 [21] and Gregory et al., 2002 [32] we found that the cluster spans in human often significantly differ from cluster spans in mouse (the span may include gaps and unaligned regions that contribute to these differences). Note that although different clusters are not supposed to overlap in 2-D, they often overlap in 1-D (i.e., their span intervals may overlap in human or mouse). Therefore, defining the cluster order for intermingled clusters should be done with caution. We compute the *center of mass* of all anchors forming the cluster and order clusters in human by the coordinates of their centers of masses. We assign the clusters numbers according to their order on the human genome. This lets us read off a cluster order in the mouse genome in terms of these labels.

Signs (orientations) of the resulting clusters are usually well-defined but in some cases are not obvious. The algorithm for sign assignments in GRIMM-Synteny and the theorem justifying this algorithm will be described elsewhere.

The number of clusters found depends on the value of the gap threshold G. Fig. 1b shows clusters in a region of the genome for the gap threshold $G = 100$ Kb. Increasing the gap threshold will typically merge some clusters; in this case, this region forms a single cluster at $G = 1$ Mb. The human and mouse genomes include some gaps and regions without anchors that may be longer than G. Such regions break a single synteny block into a few clusters. To combine such clusters into a single synteny block we define the notion of a strip. A *strip* is a sequence of consecutive signed clusters $i_1, \ldots, i_n$ in the first genome which either appear consecutively in the same way or in the reverse order $-i_n, \ldots, -i_1$ in the other genome. For example, for $G = 1$ Mb and $C = 1$ Mb, the number of clusters in the human and mouse genomes is 319 while the number of strips (synteny blocks) is 281. Most synteny blocks correspond to a single cluster but some synteny blocks contain as many as 5 clusters.

On the X chromosome, comparing Figs. 2a–b, most discarded material is very small, but there is a region near the red cluster, at human 84.6–88.6 Mb, mouse 94.3–99.7 Mb, which forms three clusters. Each has length under $C = 1$ Mb in human, so they are discarded. Increasing G or lowering C sufficiently would retain these clusters and possibly merge them with the red cluster. If they were the only addition, the red block would be larger, but the synteny block order in Fig. 2e would not be affected, so the rearrangement analysis described below would remain the same. If distinct blocks were added, it would affect the rearrangement analysis. The chosen values of G and C result in a classification of the anchor arrangements into micro-rearrangements, macro-rearrangements, and noise. Rearrangements of anchors within a synteny block are called *micro-rearrangements*. Rearrangements of the order and orientation of synteny blocks are called *macro-rearrangements*.

4 From synteny blocks to the breakpoint graph

We illustrate the notion of the breakpoint graph using the X chromosome as an example. The signed permutation describing synteny block order on the X chromosome in mouse is

$$-4,\ -5,\ 3,\ 11,\ -2,\ 8,\ -9,\ 10,\ -6,\ 7,\ -1\ .$$

For our goals, we shall use

$$1,\ -7,\ 6,\ -10,\ 9,\ -8,\ 2,\ -11,\ -3,\ 5,\ 4$$

(a "flip" of the entire chromosome). We may transform this permutation into the "identity" permutation representing the human X chromosome

$$1,\ 2,\ 3,\ 4,\ 5,\ 6,\ 7,\ 8,\ 9,\ 10,\ 11$$

by 7 reversals (Fig. 2h) via the Hannenhalli-Pevzner algorithm [10]. This algorithm uses the breakpoint graph (Fig. 2g) to construct a most parsimonious evolutionary scenario (Fig. 2h) in polynomial time. We now show a new way to construct the breakpoint graph.

Fig. 3a presents the genomic dot-plot (with added "start" and "end" elements) and the "human" path (shown with solid edges) traversing the synteny blocks in human order. The projection of this path on the human genome is shown below the human axis. Similarly, Fig. 3b presents the same genomic dot-plot and the "mouse" path (shown with dotted edges) traversing the synteny blocks in mouse order. The *2-dimensional breakpoint graph* is obtained by superimposing these solid and dotted paths (Fig. 3c) and further deleting the synteny blocks (Fig. 3d). One can prove that the breakpoint graph is a collection of alternating solid-dotted cycles. Fig. 3d consists of a cycle of length 2 (containing the start vertex), two cycles of length 4, a cycle of length 6, and a cycle of length 8. After constructing the cycles, we usually color the edges so that each cycle has its own color (Fig. 2g). At this point, we apply the Hannenhalli-Pevzner algorithm to obtain a most parsimonious scenario. Note that the breakpoint graphs in Fig. 2 and Fig. 3 are different due to the X chromosome "flip."

Hannenhalli and Pevzner, 1995 [10] demonstrated that the cycles and their interleaving structure are the "fossil records" of rearrangement events and showed how to use them for solving reversal distance and genomic distance problems. The "2-dimensional" representation of breakpoint graphs shown in Fig. 3 is different from the representation used by Hannenhalli and Pevzner, 1995 [10] (they used the "1-dimensional" projections of this graph shown along the axis in Fig. 3). We feel that the 2-dimensional breakpoint graph is a better visualization than the 1-dimensional one, in addition to being independent of the choice of the axis. Therefore, it provides better "geometrical" intuition for the Hannenhalli-Pevzner theory.

Since every reversal creates at most two new breakpoints, the reversal distance is at most half the number of the breakpoints in the genome. If there is no breakpoint re-use then the reversal distance is exactly half the number of breakpoints. Moreover, in this case the real evolutionary scenario is a most parsimonious one, thus implying that the reversal distance equals the real distance in this case. However, the estimate of reversal distance as half the number of breakpoints is inaccurate since it assumes that the breakpoints are not re-used in evolution. In most genome rearrangement studies, there are breakpoint re-uses (at least at a certain level of synteny block resolution), thus indicating that breakpoint re-use is the rule rather than the exception. For example, Palmer and Herbon, 1988 [26] describe an evolutionary scenario with one breakpoint re-use for cabbage and turnip mtDNA, while Bafna and Pevzner, 1995 [2] describe an evolutionary scenario with four breakpoint re-uses in the rearrangement scenario for human and mouse X chromosomes.

Given an evolutionary scenario, we call two breakpoints, B_1 and B_2, *siblings* if they are endpoints of a reversal R from this scenario. Two breakpoints are *related* if there is a series of breakpoints $B_1, B_2, \ldots, B_n$ such that every two consecutive breakpoints in this series are siblings. In addition, reconstructing a most parsimonious scenario and identification of related breakpoints are non-trivial tasks, even in the absence of breakpoint re-use. The breakpoint graph reveals related breakpoints and allows one to find the most parsimonious scenarios. In the scenario shown in Fig. 2h, the breakpoint at the start (flat end) of block 4 is used twice. We emphasize that by re-using breakpoints we do not mean multiple use of exactly the same genomic position as an endpoint of rearrangements, but rather the fact that between synteny blocks, there are regions that host endpoints for multiple rearrangement events.

In contrast to unichromosomal breakpoint graphs (consisting of cycles), the multichromosomal breakpoint graph consists of both cycles and paths (such paths correspond to breaking the cycles at the chromosome endpoints). The breakpoint graph for the entire human and mouse genome is very complicated (Fig. 4).

5 From the breakpoint graph to rearrangement scenarios

A previous analysis of comparative maps of human and mouse X chromosomes revealed 8 syntenic blocks and postulated a most parsimonious rearrangement scenario with 6 inversions (Bafna and Pevzner, 1995 [2]). The genomic sequences reveal 11 synteny blocks of 1 Mb and longer and provide evidence for at least 7 inversions (Fig. 2h). Moreover, there are 177 micro-rearrangements within the X chromosome that were beyond the resolution of previous comparative mapping studies (some of them may be artifacts of assembly errors). Two out of 11 synteny blocks on the X chromosome show evidence of extensive micro-rearrangements.

These estimates are based on the Hannenhalli and Pevzner, 1995 [10] theorem that expresses the reversal distance between two genomes as $n+1-c+h$, where n is the number of synteny blocks, c is the number of cycles in the breakpoint graph, and h is another easily computable combinatorial parameter of the breakpoint graph. Since h equals zero for many biological datasets, $n+1-c$ is usually a good approximation for reversal distance. For the X chromosome as depicted in Fig. 2g, we have $n = 11$, $c = 4$, $h = 0$, and we obtain a reversal distance of 8. However, flipping the whole mouse X chromosome results in the breakpoint graph of Fig. 3d, with $n = 11$, $c = 5$, $h = 0$, and reversal distance 7. Flipping a whole chromosome does not count as a rearrangement event, so the genomic distance on the X chromosome between human and mouse is 7.

A similar theorem (Hannenhalli and Pevzner, 1995b [9], Tesler, 2002 [35]) holds for multichromosomal genomes, and automatically takes into account whole chromosome flips. We used a fast implementation of the Hannenhalli-Pevzner algorithm (Tesler, 2002 [36]) to analyze the human-mouse rearrangement scenario (available via the GRIMM web server at http://www-cse.ucsd.edu/groups/bioinformatics/GRIMM/index.html). Although the algorithm finds a most parsimonious scenario, the real scenario is not necessarily a most parsimonious one (Blanchette et al., 1996 [3]), and the order of rearrangement events within a most parsimonious scenario often remains uncertain. Availability of three or more mammalian genomes could remedy some of these limitations and provide a means to infer the gene order in the mammalian ancestor (Bourque and Pevzner, 2002 [4]).

The GRIMM algorithm constructs a most parsimonious evolutionary scenario between human and mouse genomes with 245 rearrangements. With at least 245 rearrangements between human and mouse and estimated 83 Myr (Huchon et al., 2002 [12]) of evolution from their common ancestor, we obtain an estimated rate of 1.5 chromosomal rearrangements per Myr, which is higher than the previous estimate of 1.0 (Lander et al., 2001 [17]). However, this estimate should not be viewed as typical for mammalian evolution since rodents may have unusually rapid chromosome alterations. The comparative mapping data for cat and cow may soon shed further light on the comparative rates of rearrangements in different branches of the mammalian evolutionary tree.

The human-mouse breakpoint graph provides insights into rearrangements that may have occurred in the course of evolution. Some of these rearrangements are almost "obvious" (they correspond to short cycles in the breakpoint graph), while others involve long series of interacting breakpoints. Such complicated rearrangement events are described by long cycles/paths in the breakpoint graph. The longest path in the human-mouse breakpoint graph involves 26 breakpoints. The human-mouse breakpoint graph has 6 other long paths with more than 10 breakpoints.

The analysis of micro-rearrangements within the synteny blocks demonstrates a large variation in the rate of micro-rearrangements (reversals) along the genomes. In particular, 41 out of 281 synteny blocks do not show any evidence of micro-rearrangements, while 10 synteny blocks are extremely rearranged (40 or more rearrangements within a block). For example, a long synteny block on human chromosome 13/mouse chromosome 8 (nucleotides 101,902,085 to 113,413,125 on human chromosome 13) consists of 65 regions of local similarity whose order is perfectly conserved in human and mouse. On the other hand, a long synteny block on human chromosome 18/mouse chromosome 17 (positions 2,789,316 to 10,083,804 on human

chromosome 18) consists of 143 regions of local similarity and has a large number of micro-rearrangement breakpoints, indicating that there were at least 85 inversions within this block (Fig. 1b). The length of this synteny block in mouse is smaller than in human (6.0 Mb vs. 7.3 Mb). Of course, some of the breakpoints within this synteny block may be caused by assembly errors. There is evidence of at least 3170 micro-rearrangements within all the synteny blocks, some of which may be due to assembly errors.

Every breakpoint defines two synteny blocks A and B that are adjacent in one genome but separated in the second one. We distinguish between unichromosomal breakpoints (A and B belong to the same chromosome in the second genome) and multichromosomal breakpoints (A and B are on different chromosomes). Most breakpoints in the human and mouse genomes are unichromosomal breakpoints, thus indicating that most rearrangemenents that happened in the course of human-mouse evolution are intrachromosomal inversions. In particular, one can come up with a most parsimonious rearrangement scenario that includes 134 reversals in the human and mouse genomes before any translocations/fusions/fissions happen. After performing these reversals, the number of synteny blocks is reduced from 281 to 144. The breakpoint graph of these human and mouse "pre-ancestors" allows one to infer which pairs of chromosomes were involved in multiple translocations/fusions/fissions. The longest cycle in this graph involves 8 breakpoints located on 8 different chromosomes in human. The resulting rearrangement scenario from the mouse to human "pre-ancestor" has 15 inversions, 93 translocations, and 3 fissions. The complete scenario from mouse to human has 149 inversions, 93 translocations, and 3 fissions. (There are other combinations of 245 steps consistent with the breakpoint graph; this is the one we found with the most inversions.)

6 Conclusions

Molecular evolution studies are usually based on the analysis of individual genes rather than entire genomes. However, such widespread phenomena as horizontal gene transfer, differential gene loss, etc., often lead to situations when evolutionary trees for different genes tell different stories. An alternative approach is to infer the evolutionary history of the entire genomes, rather than individual genes, based on the analysis of gene orders. Although this approach is successful in bacterial genomics (see Wolf et al., 2002 [39] for a recent review), its applications in mammalian genomics are somewhat limited due to incompleteness of gene order data derived from comparative maps. Human and mouse genomic sequences, for the first time, provide a possibility to accurately estimate the extent of rearrangement events. However, the "original synteny" problem (Nadeau and Sankoff, 1997 [23]) remains unsolved since at least three mammalian gene orders are required to derive the ancestral mammalian karyotype. The ongoing mammalian sequencing projects and recently developed algorithms for reconstructing ancestral gene orders (Bourque and Pevzner, 2002 [4]) provide hope that the "original synteny" problem will finally be resolved.

7 Acknowledgements

We are grateful to Ewan Birney, Guillaume Bourque, and Bill Murphy for many helpful suggestions, and to Bernard Moret for providing his group's genome rearrangement programs. We are also indebted to Michael Kamal, Kerstin Linblad-Toh, and Jade Vinson for their advice on synteny blocks and rearrangements in human and mouse genomes.

References

[1] V. Bafna and P.A. Pevzner. Genome rearrangements and sorting by reversals. In *Proceedings of the 34th Annual IEEE Symposium on Foundations of Computer Science*, pages 148–157, 1993 (Full version has appeared in SIAM J. Computing, 25: 272-289, 1996).

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

[3] M. Blanchette, T. Kunisawa, and D. Sankoff. Parametric genome rearrangements. *Gene*, 172:GC:11–17, 1996.

[4] G. Bourque and P.A. Pevzner. Genome-scale evolution: reconstructing gene orders in the ancestral species. *Genome Research*, 12:9748–9753, 2002.

[5] E.A. Carver and L. Stubbs. Zooming in on the human-mouse comparative map: genome conservation re-examined on a high-resolution scale. *Genome Research*, 7:1123–1137, 1997.

[6] R.W. DeBry and M.F. Seldin. Human/mouse homology relationships. *Genomics*, 33:337–351, 1996.

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

[8] W. Fujibuchi, H. Ogata, H. Matsuda, and M. Kanehisa. Automatic detection of conserved gene clusters in multiple genomes by graph comparison and p-quasi grouping. *Nucleic Acids Res.*, 28:4029–4036, 2000.

[9] S. Hannenhalli and P.A. Pevzner. Transforming men into mice (polynomial algorithm for genomic distance problem). In *Proceedings of the 36th Annual IEEE Symposium on Foundations of Computer Science*, pages 581–592, Milwaukee, Wisconsin, 1995.

[10] S. Hannenhalli and P.A. Pevzner. Transforming cabbage into turnip (polynomial algorithm for sorting signed permutations by reversals). In *Proceedings of the 27th Annual ACM Symposium on the Theory of Computing*, pages 178–189, 1995 (full version appeared in Journal of ACM, 46: 1–27, 1999).

[11] R.C. Hardison, J. Oeltjen, and W. Miller. Long human-mouse sequence alignments reveal novel regulatory elements: a reason to sequence the mouse genome. *Genome Research*, 7:959–966, 1997.

[12] D. Huchon, O. Madsen, M.J. Sibbald, K. Ament, M.J. Stanhope, F. Catzeflis, W.W. de Jong, and E.J. Douzery. Rodent phylogeny and a timescale for the evolution of glires: evidence from an extensive taxon sampling using three nuclear genes. 19:1053–1065, 2002.

[13] J. Kececioglu and D. Sankoff. Exact and approximation algorithms for the inversion distance between two permutations. *Algorithmica*, 13:180–210, 1995.

[14] W. J. Kent. BLAT–the BLAST-like alignment tool. *Genome Research*, 12:656–664, 2002.

[15] B.F. Koop and L. Hood. Striking sequence similarity over almost 100 kilobases of human and mouse T-cell receptor DNA. *Nature Genetics*, 7:48–53, 1994.

[16] S. Kumar, S.R. Gadagkar, A. Filipski, and X. Gu. Determination of the number of conserved chromosomal segments between species. *Genetics*, 157:1387–1395, 2001.

[17] E. Lander and et al. Initial sequencing and analysis of the human genome. *Nature*, 409:860–921, 2001.

[18] W.C. Lathe, B. Snel, and P. Bork. Gene context conservation of a higher order than operons. *Trends Biochem Sci.*, 25:474–479, 2000.

[19] B. Ma, J. Tromp, and M. Li. PatternHunter: faster and more sensitive homology search. *Bioinformatics*, 18:440–445, 2002.

[20] C. Mayor, M. Brudno, J.R. Schwartz, A. Poliakov, E. M. Rubin, K. A. Frazer, L. Pachter, and I. Dubchak. VISTA: Visualizing global DNA sequence alignments of arbitrary length. *Bioinformatics*, 16:1046–1047, 2000.

[21] R.J. Mural and et al. A comparison of whole-genome shotgun-derived mouse chromosome 16 and the human genome. *Science*, 296:1661–1671, 2002.

[22] N. Copeland, et al. A genetic linkage map of the mouse: Current applications and future prospects. *Science*, 262:57–66, 1993.

[23] J.H. Nadeau and D. Sankoff. Landmarks in the Rosetta Stone of mammalian comparative maps. *Nature Genetics*, 15:6–7, 1997.

[24] 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.

[25] S.J. O'Brien, M. Menotti-Raymond, W.J. Murphy, W.G. Nash, J. Wienberg, R. Stanyon, N.G. Copeland, N.A. Jenkins, J.E. Womack, and J.A. Graves. The promise of comparative genomics in mammals. *Science*, 286:458–481, 1999.

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

[27] R. Puttagunta, L.A. Gordon, G.E. Meyer, D. Kapfhamer, J.E. Lamerdin, P. Kantheti, K.M. Portman, W.K. Chung, D.E. Jenne, A.S. Olsen, and M. Burmeister. Comparative maps of human 19p13.3 and mouse chromosome 10 allow identification of sequences at evolutionary breakpoints. *Genome Research*, 10:1369–1380, 2000.

[28] R. Waterston, et al. Initial sequencing and comparative analysis of the mouse genome. *Nature*, 420:520–562, 2002.

[29] I.B. Rogozin, K.S. Makarova, J. Murvai, E. Czabarka, Y.I. Wolf, R.L. Tatusov, L.A. Szekely, and E.V. Koonin. Connected gene neighborhoods in prokaryotic genomes. *Nucleic Acids Res*, 30:2212–2223, 2002.

[30] D. Sankoff and M. Blanchette. The median problem for breakpoints in comparative genomics. In *Computing and Combinatorics, Proceeedings of COCOON '97*, Lecture Notes in Computer Science, pages 251–263, New York, 1997. Springer Verlag.

[31] S. Schwartz, Z. Zhang, K. A. Frazer, A. Smit, C. Riemer, J. Bouck, R. Gibbs, R. Hardison, and W. Miller. PipMaker – a web server for aligning two genomic DNA sequences. *Genome Research*, 10:577–586, 2000.

[32] et al. S.G. Gregory. A physical map of the mouse genome. *Nature*, 418:743–750, 2002.

[33] A.H. Sturtevant and T. Dobzhansky. Inversions in the third chromosome of wild races of *Drosophila pseudoobscura*, and their use in the study of the history of the species. *Proceedings of the National Academy of Sciences USA*, 22:448–450, 1936.

[34] R.L. Tatusov, E.V. Koonin, and D.J. Lipman. A genomic perspective on protein families. *Science*, 278:631–637, 1997.

[35] G. Tesler. Efficient algorithms for multichromosomal genome rearrangements. *J. Comp. Sys. Sci.*, 65(3):587–609, 2002.

[36] G. Tesler. GRIMM: genome rearrangements web server. *Bioinformatics*, 18:492–493, 2002.

[37] J.W. Thomas, T.J. Summers, S.Q. Lee-Lin, V.V. Maduro, J.R. Idol, S.D. Mastrian, J.F. Ryan JF, D.C. Jamison, and E.D. Green. Comparative genome mapping in the sequence-based era: early experience with human chromosome 7. *Genome Research*, 10:624–633, 2000.

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

[39] Y. Wolf, I. Rogozin, N. Grishin, and E. Koonin E. Genome trees and the tree of life. *Trends Genet*, 18:472, 2002.

[40] Y.I. Wolf, I.B. Rogozin, A.S. Kondrashov, and E.V. Koonin. Genome alignment, evolution of prokaryotic genome organization, and prediction of gene function using genomic context. *Genome Research*, 11:356–372, 2001.

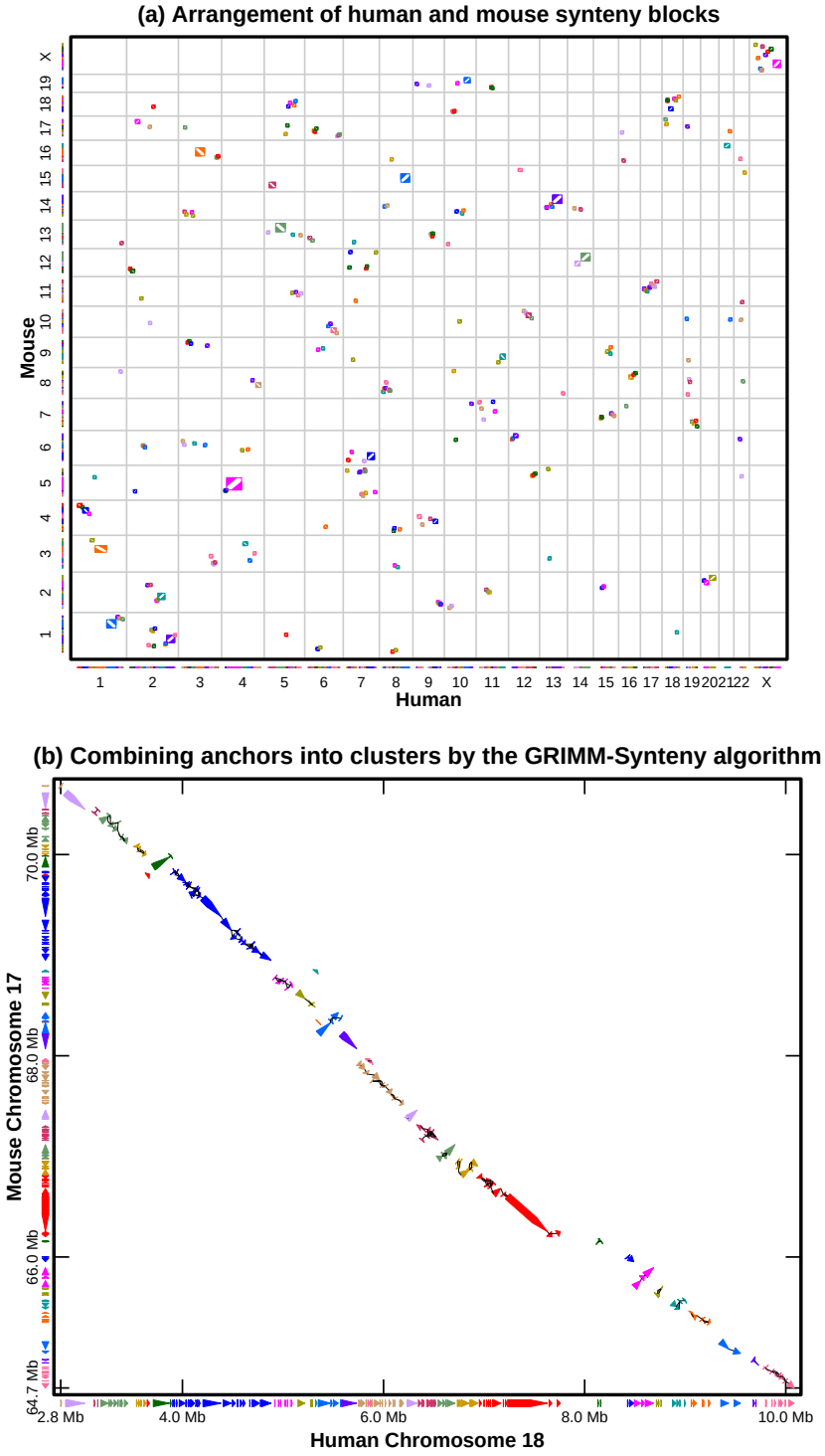

Figure 1: (a) Human and mouse synteny blocks. Every block corresponds to a rectangle, with a diagonal showing whether the arrangements of anchors in human and mouse (within the synteny block) are the same or reversed. (b) Combining anchors into clusters by the GRIMM-Synteny algorithm at $G = 100$ Kb. The edges in the anchor graph connect the closest ends of the anchors. The anchors are color-coded by the resulting clusters. At $G = 1$ Mb, this forms a single cluster, which in turn forms a synteny block (the lower right block in the human 18/mouse 17 rectangle in (a)).

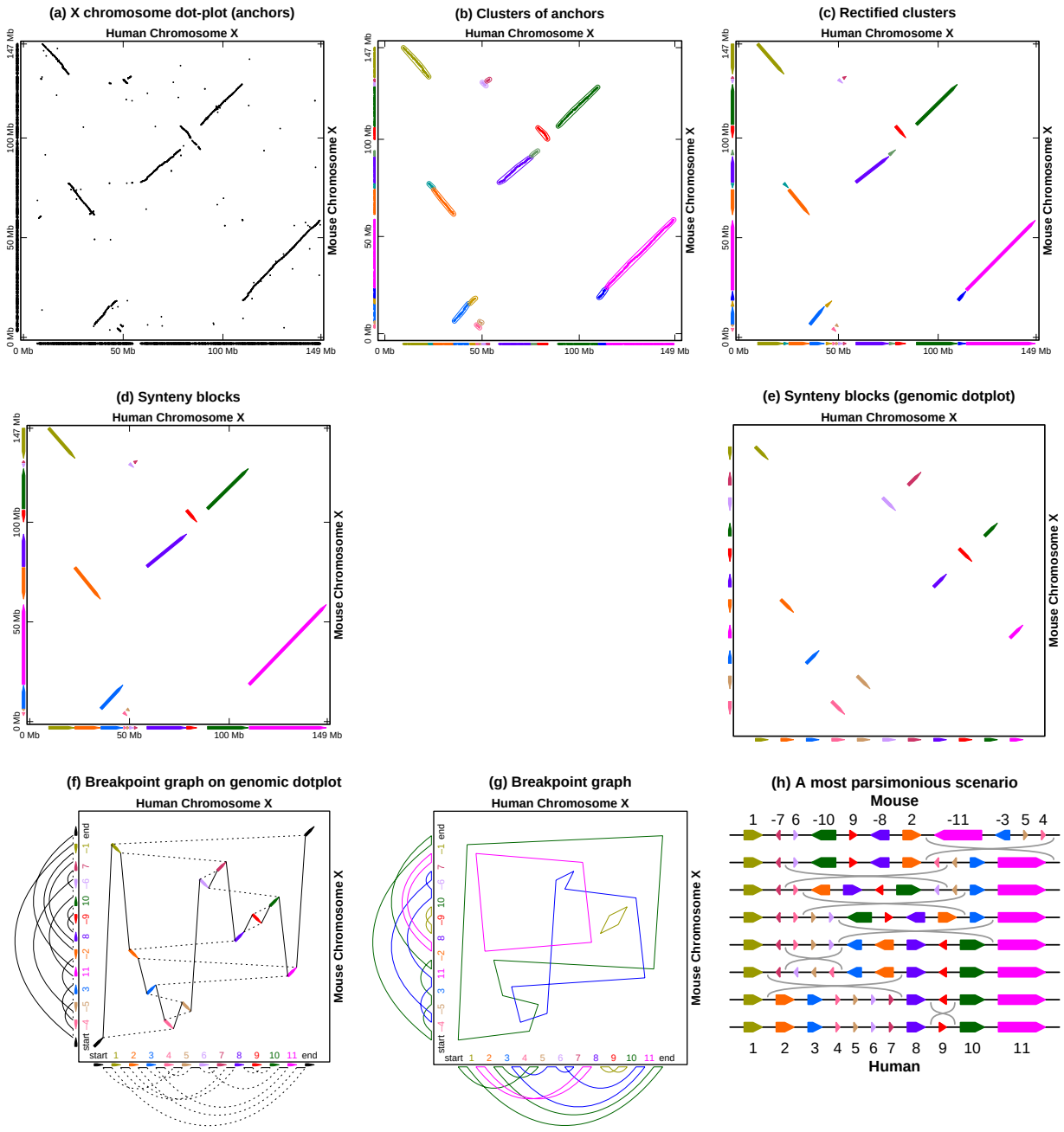

Figure 2: X chromosome: from local similarities, to synteny blocks, to breakpoint graph, to rearrangement scenario. (a) Dot-plot of anchors. Anchors are enlarged for visibility. (b) Clusters of anchors. (c) Rectified clusters. (d) Synteny blocks. (e) Synteny blocks (symbolic representation as genome rearrangement units). (f) 2-D breakpoint graph superimposed on synteny blocks. The projections of the 2-D graph onto the human and mouse axes form the conventional breakpoint graphs. (g) 2-D breakpoint graph. The four cycles in the breakpoint graph are shown by different colors. (h) A most parsimonious rearrangement scenario for human and mouse X chromosomes.

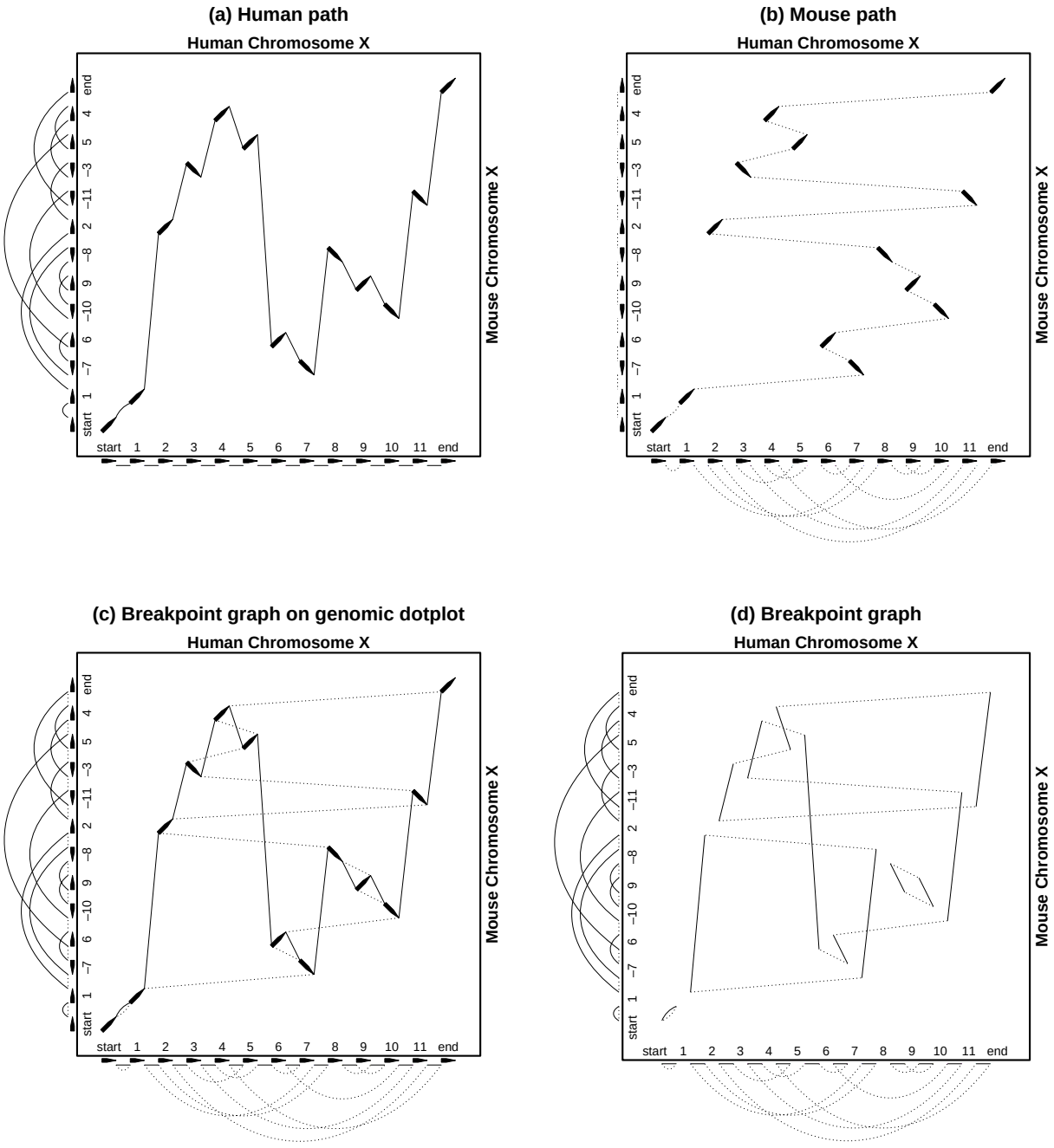

Figure 3: Construction of the breakpoint graph from synteny blocks. (a) Solid path through human. (b) Dotted path through mouse. (c) Superposition of paths. (d) Remove blocks to obtain cycles.

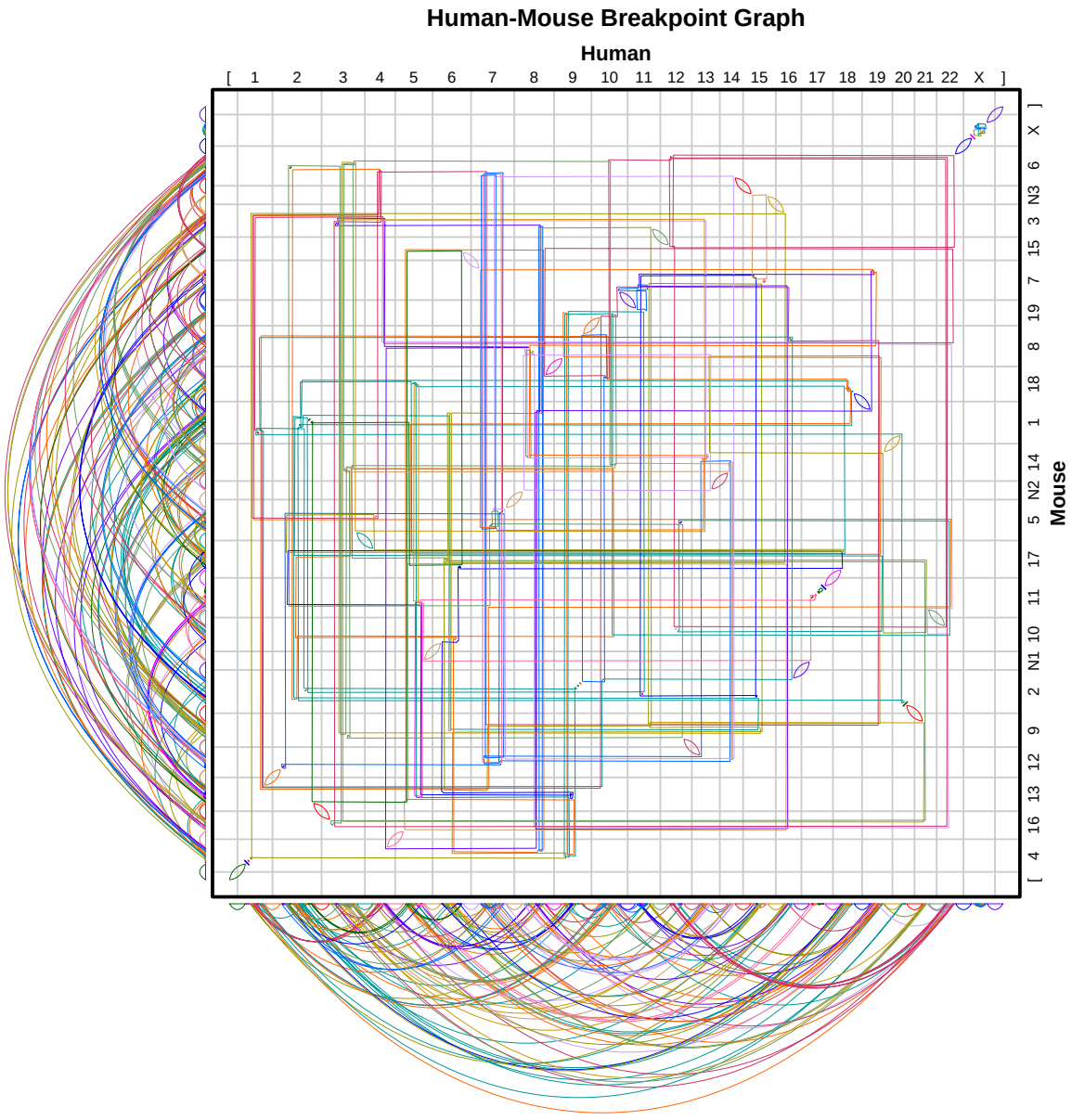

Figure 4: Multichromosomal breakpoint graph of the whole human and mouse genomes. The conventional chromosome order and orientation is not suitable for such graphs; an optimal chromosome order and orientation was determined by the algorithm in Tesler, 2002 [35]. Three "null chromosomes" N1, N2, N3, were added to mouse to equalize the number of chromosomes in the two genomes.