Sankoff D, Zheng C, Muñoz A *et al.* Issues in the reconstruction of gene order evolution. JOURNAL OF COMPUTER SCIENCE AND TECHNOLOGY 25(1): 10–25 Jan. 2010

Issues in the Reconstruction of Gene Order Evolution

David Sankoff[1], Chunfang Zheng[2,3] (郑春芳), Adriana Muñoz[4], Zhenyu Yang[1] (杨振宇), Zaky Adam[4,5], Robert Warren[4], Vicky Choi[6] (蔡维仁), and Qian Zhu[7,8] (祝 谦)

[1] *Department of Mathematics and Statistics, University of Ottawa, Ottawa, K1N 6N5, Canada*

[2] *Department of Biology, University of Ottawa, Ottawa, K1N 6N5, Canada*

[3] *Département d'informatique et de recherche opérationnelle, Université de Montréal, H3C 3J7, Canada*

[4] *School of Information Technology and Engineering, University of Ottawa, Ottawa, K1N 6N5, Canada*

[5] *Biology Department, The Pennsylvania State University, University Park, PA 16802, U.S.A.*

[6] *Department of Computer Science, Virginia Polytechnic Institute, Falls Church, VA 22043, U.S.A.*

[7] *Department of Biochemistry, University of Ottawa, Ottawa, K1N 6N5, Canada*

[8] *Department of Computer Science, Princeton University, Princeton, NJ 08544, U.S.A.*

E-mail: {sankoff, amun010, zyang009, zadam008, rwarr059}@uottawa.ca; chunfang313@gmail.com; vchoi@cs.vt.edu; qzhu@princeton.edu

Received September 1, 2009; revised December 14, 2009.

Abstract As genomes evolve over hundreds of millions years, the chromosomes become rearranged, with segments of some chromosomes inverted, while other chromosomes reciprocally exchange chunks from their ends. These rearrangements lead to the scrambling of the elements of one genome with respect to another descended from a common ancestor. Multidisciplinary work undertakes to mathematically model these processes and to develop statistical analyses and mathematical algorithms to understand the scrambling in the chromosomes of two or more related genomes. A major focus is the reconstruction of the gene order of the ancestral genomes.

Keywords chromosome rearrangement, comparative genomics, gene clusters, phylogenomics

1 Introduction

The advent of nuclear genome sequencing in eukaryotes, especially in mammals[1-2], lent great impetus to the field of comparative structural genomics. During the previous decade, computer scientists had been preparing for this era by developing analyses and algorithms for comparing whole genomes, illustrating them on smaller genomes such as those of mitochondria[3-4], chloroplasts[5] and prokaryotes[6], and using various sets of ordered markers, usually genes, as the input data, rather than raw sequence.

The central result during this period was the polynomial-time Hannenhalli-Pevzner algorithm[7] for sorting, in a minimum number of steps, a signed genome (i.e., a signed permutation of $1, 2, \ldots, n$) into the identity permutation by successively reversing contiguous fragments of the permutation, where each reversal also switches the polarity of each term in the fragment. This allowed the rapid calculation of the genomic distance d between two genomes, since a reversal models the major chromosomal rearrangement process in biology, inversion. At the same time, Caprara[8] showed that most efficient sorting by reversals of unsigned permutations is an NP-hard problem. These results responded to a question that had been raised in various forms in the 1980s[9-10] and even earlier[11]. Extensions of the Hannenhalli-Pevzner approach[12-13] allowed the comparison of multi-chromosomal genomes, by modelling the biological processes of reciprocal translocation as well as chromosome fusion and fission.

During this time as well, the first methods of reconstructing ancestral gene orders were proposed. Using the breakpoint metric b, essentially the number of adjacent genes in one genome not adjacent, or at least not with the same sign, in the other, it became possible to infer the optimal ancestral genomes in a given phylogeny[14] for moderate size genomes. And using the genomic distance d, El-Mabrouk devised an exact polynomial-time algorithm to reconstruct the ancestral form of present-day descendants of a tetraploidization, or whole genome doubling, event[15-16].

Regular Paper

This work was supported in part by grants and fellowships from the Natural Science and Engineering Council of Canada (NSERC).

One task that becomes disproportionately more difficult in comparing nuclear genomes, with typically 3×10^9 DNA base pairs in mammals, than with mitochondrial ($10^4 \sim 10^5$ base pairs) or chloroplast (typically $1 \sim 3 \times 10^5$) genomes is that of identifying the corresponding elements (homologs, orthologs, ancestrally related) to compare. Not only is it extremely difficult and tedious just to identify the genes in a genome, and to distinguish which of potentially many approximate copies of a gene in one genome corresponds to which in the other, but most of the genome is not occupied by genes at all.

There are two types of response to this difficulty. One is to try to systematically identify conserved segments: Homologous lengths of sequence in the two genomes, using alignment techniques and without regard to the genetics or function (i.e., whether the segment contains genes, parts of genes, or regulatory elements), and then to treat these segments as ordered markers in the comparative analysis. This was first done in 2003 by Pevzner and Tesler[17] focusing on highly conserved (anchor) regions of sequence, attaining thresholds of length and similarity. At the same time Kent *et al.*[18], in building the UCSC genome browser[19], applied a new, nested alignment procedure over all regions of the genome sequence, which could then be analysed at various levels of resolution. Unfortunately, sequence-based approaches to comparative structural genomics not firmly anchored in genetic correspondences are not at all robust to changes in threshold or resolution parameters[20-21].

The second way of responding to the difficulty in homology identification is simply to wait for the result of genome annotation, i.e., the systematic expert identification of genes and other sequence elements in all the genomes being compared. This task may take years to complete, a frustratingly long delay for those eager to do comparisons, but recent advances in automated or at least computer-assisted gene finding and annotation mean that relatively accurate gene identification is increasingly available and that homologies can be quickly established, enabling genomic distance calculations soon after sequence assembly.

The advances and challenges described in this paper will be phrased in terms of the second approach. We will characterize genomes as gene orders, and often use the terms interchangeably. We will not delve further into the separate question of identifying genes within a genome, though we will have to address aspects of the question of how to treat spatially dispersed multicopy or paralogous genes in one genome with respect to their homologous counterparts in another genome, a problem which is discussed in depth by Tao Jiang elsewhere in this issue[22].

The mathematical details of some of the work discussed here appears in [23], while the field of combinatorics and algorithms for genome rearrangement is the topic of an excellent survey volume by Fertin *et al.*[24] In the next section, we sketch various kinds of genomes and how they are formalized as well as the basic chromosomal rearrangement processes and how they are defined mathematically. This leads to various concepts of genomic distance.

In Section 3 we discuss the so-called median problem as the basis for the reconstruction of genomes at the ancestral nodes of phylogenetic trees and in Section 4 we reconstruct the diploid ancestor whose tetraploidization (whole genome duplication) is apparent in the gene complement of its present-day descendants. Then we introduce guided genome halving, which improves the accuracy of this reconstruction. Moreover, where gene order phylogenies have been hitherto been confined to diploid genomes, guided halving enables polyploid genomes to be incorporated in a principled way into these phylogenies.

In Section 5, we examine various ways in which the available data may fall short of the usual requirements of complete and accurate linear gene orders along each chromosome. For certain types of data only partial ordering of the genes is available, and we discuss how to linearize the corresponding DAG (directed acyclic graph) through genome comparison. Another problem is due to incompletely assembled genomes, where the linear order on a chromosome may be fragmented into many smaller linear orders called contigs. We examine the consequences for gene order phylogenetics of treating each of these contigs as if it were a complete chromosome. Finally, gene order data may contain errors, so that a proportion of genes are in the wrong place on a chromosome or even on the wrong chromosome. We show a way of detecting these errors and purifying the data through genome comparison.

Some genomes may be so thoroughly scrambled during evolution that it may not be meaningful to attempt to reconstruct the entire rearrangement history separating two genomes. This may be particularly true of some prokaryotic comparisons. Instead we search for groups of genes which are significantly closer to each other in both genomes than would be the case by chance. Section 6 studies ways of defining, detecting and testing these gene clusters.

2 Genomes, Distances and Evolutionary Time

In this section, we first will sketch various formalizations of the biological concepts of chromosome

and genome, more specifically the mathematical abstractions of genome content, chromosome shape, gene sign and gene order. We introduce comparative genomics in terms of *breakpoints* and the *breakpoint distance*. Then we will define evolutionary operations that affect genomes and discuss how different combinations of genome structure and permitted evolutionary operations result in computationally different models.

We introduce genomic rearrangement distances and explain how they apply to different kinds of genomes. We then explore the dependance of expected genomic distance on time, or rather on the number of evolutionary steps under the constant rate-of-change hypothesis. As genomes diverge over time, measures evaluating the differences between these genomes increase. Even in the simplest model, this increase cannot be linear indefinitely over time, since any such measure will have a maximum value, typically the expected difference between two randomly permuted genomes.

2.1 Genomes

We start with a set G of distinct elements called *genes*. A gene g is represented by its two *ends*, its *head* g_h and a *tail* g_t. (In biochemical terms, the heads may correspond to the $5'$ ends of the genes and the tails the $3'$ ends, or vice-versa.) An *adjacency* is an unordered pair of gene ends, generally but not necessarily from different genes; a *genome* is a set of adjacencies on G, where no gene end is in more than one adjacency. A gene end that is not in any adjacency is called a *telomere*. Consider a gene g, together with any gene h having an end adjacent to an end of g (there may be 2, 1 or 0 such h), together with any other gene k having an end adjacent to the other end of h, and so on, accumulating genes by transitivity of adjacency. The subset of G thus constructed is called a *chromosome*. If a chromosome contains two telomeres it is a linear chromosome, if it contains no telomere it is circular. A genome with only linear chromosomes is called a *linear* genome.

A genome with only one chromosome is called *unichromosomal*; if it has more than one chromosome it is *multichromosomal*.

As well as the representation in terms of sets of adjacencies, a genome can also be represented as a set of strings, by writing the genes for each chromosome starting with one that has a telomeric end and adding genes according to the adjacencies of their ends. Each gene g whose tail is written first is considered to have positive polarity, while a gene whose head is written first has negative polarity ($-g$). In this way, unichromosomal genomes are equivalent to *signed permutations* by virtue of the head-tail polarity of the gene ends, irrespective of whether they are linear or circular. For each linear chromosome, there are two possible equivalent strings, according to the arbitrary chosen telomere. One string is obtained from the other by reversing the order and switching the signs of all the genes. For circular chromosomes, there are also two possible circular string representations, according to the direction in which the genes are traversed.

Although we have formulated genomes in terms of sets of distinct genes, in biological reality there are often many copies, identical or almost so, of the same gene in a genome. Incorporating this fact into the mathematics of genome comparison and genome reconstruction complicates the formulation of problems and inevitably worsens their complexity. For this introductory essay, therefore, we will keep largely to the single-copy genes case, and leave it to the reader to explore the voluminous literature (starting with [22]) that attempts to generalize to the multicopy case. We will touch on the latter occasionally, especially in Section 4 on genome halving, a context where there are exactly two copies of each gene.

2.2 Breakpoints

For two genomes on the same set G containing n genes, suppose $\{x, y\}$ is an adjacency in one of the genomes but not the other. This is called a *breakpoint*.

Let a be the number of adjacencies in common in the two genomes, and e be the number of telomeres in common. Then the *breakpoint distance* is

$$d_{\mathrm{BP}}(\Pi, \Gamma) = n - a - \frac{e}{2}. \tag{1}$$

This definition from [23], applies to the comparison of all the types of genomes mentioned in Subsection 2.1. It may differ from other definitions slightly, largely in terms of how it accounts for differing sets of telomeres.

2.3 Operations

The classical genetics notions of inversion, transposition and reciprocal translocation of chromosome segments, as well as chromosomal fission and fusion, are formalized in such papers as those by Tesler[13], Yancopoulos *et al.*[25], and Bergeron *et al.*[26] Briefly, using the string representation of a chromosome, e.g., $h_1 \cdots h_l$, where a pair of successive genes $h_u h_{u+1}$ are loosely termed an adjacency if their gene ends constitute an adjacency, we can illustrate:

• an inversion (implying change of sign, i.e., change of strand) of a chromosomal segment:

$$h_1 \cdots \underline{h_u \cdots h_v} \cdots h_m \rightarrow h_1 \cdots \underline{-h_v \cdots - h_u} \cdots h_m,$$

disrupting the two adjacencies $h_{u-1}h_u$ and $h_v h_{v+1}$,

• a transposition of a chromosomal segment:

$$h_1 \cdots h_u \cdots \underline{h_v \cdots h_w} \cdots h_m \rightarrow$$
$$h_1 \cdots h_{u-1} \underline{h_v \cdots h_w} h_u \cdots h_{v-1} h_{w+1} \cdots h_m,$$

disrupting the three adjacencies $h_{u-1}h_u$, $h_{v-1}h_v$ and $h_w h_{w+1}$,

• a reciprocal translocation between two chromosomes:

$$h_1 \cdots \underline{h_u \cdots h_l}, k_1 \cdots \underline{k_v \cdots k_m} \rightarrow$$
$$h_1 \cdots \underline{k_v \cdots k_m}, k_1 \cdots \underline{h_u \cdots h_l},$$

disrupting the two adjacencies $h_{u-1}h_u$ and $k_{v-1}k_v$,

• a chromosome fission:

$$h_1 \cdots \underline{h_v h_{v+1}} \cdots h_l \rightarrow h_1 \cdots h_v, h_{v+1} \cdots h_l,$$

disrupting the adjacency $h_v h_{v+1}$, and

• the fusion of two chromosomes:

$$h_1 \cdots h_l, k_1 \cdots k_m \rightarrow h_1 \cdots \underline{h_l k_1} \cdots k_m.$$

The genomic distance is the minimum number of operations of these types (or some specified subset of types) required to transform one of the genomes being compared into the other. The authors mentioned above also provide rapid algorithms for deriving the distance, given genomes composed of ordered chromosomes represented by the same n genes, markers or segments in the two genomes, assuming the strandedness, or reading direction, of each gene is known.

2.4 Rearrangement Distance

A double-cut-and-join (DCJ) is an operation acting on two adjacencies $\{p, q\}$ and $\{r, s\}$, deleting them and replacing them by $\{p, r\}$ and $\{q, s\}$ or by $\{p, s\}$ and $\{q, r\}$. Also it can act on an adjacency $\{p, q\}$ and a telomere r to produce the adjacency $\{p, r\}$ and a telomere q or the adjacency $\{q, r\}$ and a telomere p. It can also fuse two telomeres to create an adjacency or fission an adjacency to create two telomeres.

A DCJ can have the effect of inverting an interval of a genome, fission one chromosome into two, fusing two chromosomes into a single one, or producing a reciprocal translocation between two chromosomes. Two consecutive DCJ operations may result in a block interchange: two arbitrary segments of the genome exchange their positions, a particular case is that of a transposition, for which the two segments are contiguous. The DCJ operation is thus a very general framework. It was introduced by Yancopoulos *et al.*[25] and was simplified by Bergeron *et al.*[26]

The minimum number of DCJ operations needed to transform one genome into another (on the same set of genes) is the *DCJ distance* d_{DCJ}. It can be quickly calculated by defining a bipartite graph where the adjacencies and telomeres of one genome constitute the vertices on one side of the graph and the adjacencies and telomeres of the other genome are the opposing set of vertices. An edge is drawn between two vertices (from the two genomes) if they are both derived from adjacencies or telomeres containing a same gene end. Then, it is proved in [26],

$$d_{\mathrm{DCJ}} = n - c - \frac{i}{2} \tag{2}$$

where c is the number of cycles in the graph and i is the number of paths with an odd number of edges.

The reversal/translocation distance was introduced by Hannenhalli and Pevzner[12], and is equivalent to the DCJ distance constrained to linear genomes.

For a linear genome, a *linear* DCJ operation is a DCJ operation that results in a linear genome. This allows reversals, chromosome fusions, fissions, and reciprocal translocations. Other DCJs that create temporary circular chromosomes, are not allowed, so that transpositions (and other block interchanges) may require three operations instead of two. Chromosomes fusions and fissions are particular cases of translocations in this framework. We call the minimum number of linear DCJ operations that transform one linear genome into another *RT distance* and we denote it by d_{RT}. This distance can be calculated rapidly with formulae analogous to (2), but which have rather complex form[12-13].

2.5 Relation Between True and Inferred Distance

Even assuming that rearrangements occur at a relatively constant rate over time and are randomly positioned in the genomes, we have no simple, exact probability relationship between the actual number τ of rearrangements after a certain time t has elapsed and the number of rearrangements d inferred by applying the genomic distance algorithms to compare the initial and the derived genomes[27-29]. We can, however, model the proportion of adjacencies that will be disrupted versus the proportion that will remain intact after τ random rearrangements. For each of the adjacencies in the original genome, the probability that it will remain undisrupted after τ rearrangements is $(1 - \lambda/n)^\tau$ or approximately $e^{-\lambda\tau/n}$, where λ depends on the proportions of the various kinds of rearrangements in the model. Thus the number of disrupted adjacencies, the breakpoint distance, will be approximately $n(1 - e^{-\lambda\tau/n})$.

In analogy to the closely related breakpoint distance,

we suggested[30] that as a first approximation

$$\mathrm{E}\Big(\frac{d}{n}\Big) \approx 1 - e^{-\lambda\tau/n}, \tag{3}$$

where n is the number of genes in both genomes, and λ is a constant. If we knew λ, we could estimate τ using

$$\hat{\tau} = -\frac{n}{\lambda}\log\Big(1 - \frac{d}{n}\Big). \tag{4}$$

In fact, the relationship between the actual and inferred numbers of rearrangements (not shown) deviates considerably from the one-parameter model in (3) for both small and large τ. We thus add a quadratic correction to the linear term in the exponent, so that the model becomes

$$\mathrm{E}\Big(\frac{d}{n}\Big) \approx 1 - e^{-\lambda_1\tau/n - \lambda_2(\tau/n)^2}, \tag{5}$$

in which case the estimate of τ becomes

$$\hat{\tau} = \frac{n}{2\lambda_2}\Big(-\lambda_1 + \sqrt{\lambda_1^2 - 4\lambda_2\log\Big(1 - \frac{d}{n}\Big)}\Big). \tag{6}$$

To estimate the parameters λ_1 and λ_2, we simulated 100 pairs of genomes with almost 9000 genes and τ up to 9000 random rearrangements to derive one genome from the other. We then used a DCJ algorithm[26] to obtain an estimate of $\mathrm{E}(d)$ from the genomes.

Fig.1 shows the relationship between τ and both $\mathrm{E}(d)$ and $\hat{\tau}$, using the values $\lambda_1 = 0.846$ and $\lambda_2 = 0.576$, found by a least sum of squares criterion applied to the set of τ and $\hat{\tau}$ values. The way τ and d are normalized means that the parameters should not be very sensitive to n.

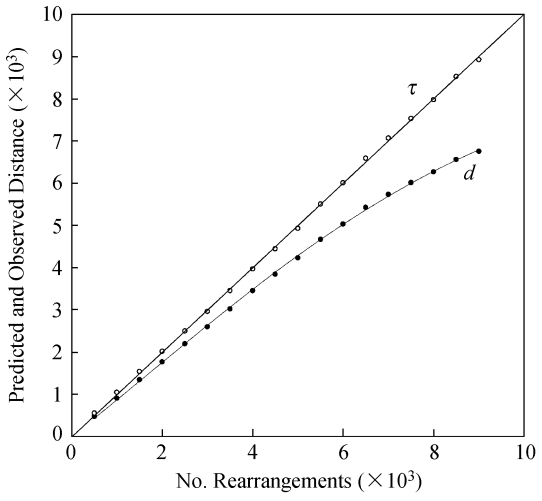


Fig.1. Predicted (curve) and observed (dots) values of genomic distance d, and inferred (open dots) values of $\hat{\tau}$ versus true (diagonal line) values.

We found that for the large values of n we studied, the model in (5) fits the simulated data, including for large and small values of τ, better than the purely curve-fitting model in [29], which also has two parameters.

3 The Median and the Small Phylogeny Problem

The evolution of species, especially eukaryotic species, is most often represented by a phylogenetic tree. The problem of reconstructing or inferring a tree from data on present-day species may be conceptually (and often methodologically) separated into two parts. The *large* phylogenetic problem is one of finding the topology, or branching pattern, of the tree connecting the given species represented by the leaves, or terminal nodes, of the tree. The *small* phylogenetic problem is the inference, for a given tree, of the ancestral species identified with each of the non-terminal nodes of the tree. In this section we will deal with the small problem in the case where the data on the present-day species are the orders of the genes on their chromosomes.

There are a variety of aspects of gene order that can be rapidly reconstructed, e.g., in [31-33], while monitoring the linearity of the reconstructed chromosomes through maintenance of bandwidth or a PQ-tree structure.

Here we will concentrate on solving the global problem by minimizing total branch length over a phylogeny while reconstructing optimal ancestral gene orders. This approach has been recently proven to be superior to local techniques when confronted with a manually validated ancestor gene order[33]. Formally, let $\mathcal{P}$ be a phylogeny where each of the N_t terminal nodes is labelled by a known gene order on the same n genes, and let d be a metric on the set of gene orders. Each branch of $\mathcal{P}$ may be incident to at most one terminal node and at least one of the N_a ancestral nodes. Each non-terminal node is of degree at least three. We want to reconstruct $R = (G_1, \ldots, G_{N_a})$, a set of gene orders at the ancestral nodes that minimize

$$L(R) = \sum_{\text{branch } XY \in \mathcal{P}} d(XY). \tag{7}$$

We use a hill-climbing procedure to find a local optimum for $L(R)$. This is illustrated in Fig.2. The archetypical (unrooted) phylogeny has three or more leaves and exactly one non-terminal node, as shown in Fig.2(a). The problem becomes that of reconstructing a single gene order M, the sum of whose distances to the given gene orders is minimal. This problem has a relatively long history, with an early algorithm[34] for the breakpoint median based on d_{BP}. Technical speedups

were described by Cosner *et al.*[5] and incorporated into the GRAPPA software[35]. Siepel[36] and Caprara[37] gave exact median algorithms for small instances of d_{RT} and Bourque[38] released a heuristic web application for this version of the problem. Much progress has been made recently on exact algorithms capable of handling large or moderate size genomes[39-40] for d_{DCJ}.

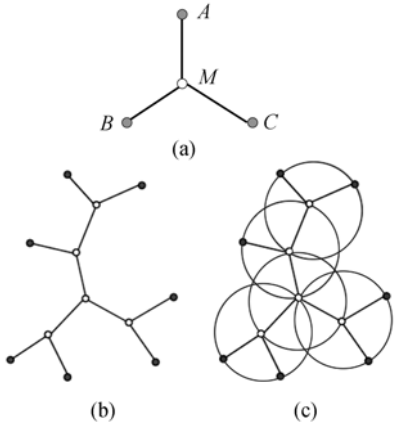


Fig.2. (a) Median problem: Given genomes A, B, C, find M such that $d(A, M) + d(B, M) + d(C, M)$ is minimized. (b) Example of unrooted phylogeny with given present-day genomes at terminal nodes (dark dots) and genomes to be inferred at the ancestral nodes (white dots). (c) Inference of genomes at ancestral nodes found by iterating through the ancestral vertices, solving a median problem at each step.

For most formulations, in terms of different kinds of genome and different distances, the median problem is known (or thought) to be NP-hard; recently, however, for the case of breakpoint distance on multichromosomal genomes not restricted to be linear, Tannier *et al.*[23] have given a polynomial-time algorithm, and this has been implemented[41] as a rapidly executing program.

Focusing on the more general small phylogeny problem with more than one ancestral node, the current heuristic strategy is based on the ability of the median algorithm to achieve a fairly accurate solution in a reasonable time on a large proportion of instances. As illustrated in Fig.2, the phylogeny at Fig.2(b) is decomposed on Fig.2(c) into a set of overlapping median configurations, with one non-terminal, i.e., ancestral, node as median, and all its (three or more) co-linear nodes, terminal or non-terminal. The heuristic consists of solving each of the median problems in turn, updating the median at each step only if it diminishes the sum of the lengths of the branches incident to the median, and iterating. This eventually converges to a local minimum. The quality of the solutions may depend on the initialization of the ancestral gene orders[14], e.g., by random gene orders, or by copying some of the present-day gene orders to the ancestral nodes. It may also depend on various techniques for escaping from local minima[42].

4 Genome Halving

4.1 Whole Genome Duplication and Halving

Many genomes have been shown to result from an ancestral doubling of the genome, often called WGD for "whole genome duplication", so that every chromosome, and hence every gene, in the entire genome is duplicated simultaneously. Evidence for the effects of genome duplication has shown up across the eukaryote spectrum, from protists, to yeast, fish, amphibians and even mammals. Genome duplication is particularly prevalent in plants.

Following WGD, rearrangements disrupt the gene order and may transfer the genes from any one chromosome onto other chromosomes. Eventually the chromosomal neighbourhood of a gene need bear no resemblance to that of its duplicate. The present-day genome can be decomposed into a set of originally duplicate genes dispersed among the chromosomes. Comparisons of genomes consisting of duplicate genes is a special case of the more general problem of genome rearrangement algorithms allowing paralogy.

The genome halving problem asks, given a genome T with two copies of each gene, distributed in any manner among the chromosomes, to find the "ancestral" genome, written $A \oplus A$, consisting of two identical halves, i.e., two identical sets of chromosomes with one copy of each gene in each half, such that the rearrangement distance $d(T, A \oplus A)$ between T and $A \oplus A$ is minimal. Note that part of this problem is to find an optimal labelling as "1" or "2" of the two genes in a pair of copies, so that all n copies labelled "1" are in one half of $A \oplus A$ and all those labelled "2" are in the other half. The genome A represents the ancestral genome at the moment immediately preceding the WGD event giving rise to $A \oplus A$.

For d_{RT}, a linear-time solution has been available for some time[15-16]. This has been adapted for d_{DCJ}[43-44], an application of DCJ that showcases its simplicity and elegance as a measure of genomic divergence. A rapid halving algorithm is also available for d_{BP}[23].

Between the computationally complex problems of gene order comparison allowing arbitrarily many copies of each gene[45], and the computationally tractable genome halving, lies the biologically plausible problem of partial genome halving. Here only part of the genome, e.g., one or several chromosomes, has been doubled. Whether some formulation of partial genome

halving could be solved efficiently is an interesting open problem.

4.2 Guided Halving

A problem with solutions to the genome halving problem is that it usually has many, very diverse, solutions. For biological purposes it would be preferable to be able to use some additional, or external, information to choose amongst these solutions. Thus the guided genome halving problem[46] asks, given T as well as another genome R containing only one copy of each of the n genes, find A so that $d(T, A \oplus A) + d(A, R)$ is minimal. The solution A need not be a solution to the original halving problem.

Nevertheless, the solution of the guided halving problem is often a solution of the original halving problem as well, or within a few rearrangements of such a solution[46-49]. This has led us to define a *constrained* version of the guided halving problem, namely to find A so that $A \oplus A$ is a solution to the original halving problem and $d(T, A \oplus A) + d(A, R)$ is minimal. This has the advantage that a good proportion of the computation, namely the halving aspect, is guaranteed to be rapid and exact, although the overall algorithm, which is essentially a search among all optimal A, remains heuristic.

4.3 Genome Aliquoting

Whole genome doubling is not the only process that results in multiple copies of each chromosome in a genome. Hexaploidy, octaploidy, etc., are conditions where the genome has been tripled, quadrupled, etc. Warren has generalized the genome halving problem to one of genome *aliquoting*[50]:

Given a genome T with $p \geqslant 2$ copies of each gene, distributed in any manner among the chromosomes, to find the "ancestral" genome, written $A \oplus A \oplus \ldots \oplus A$, consisting of p identical parts, i.e., p identical sets of chromosomes with one copy of each gene in each part, such that the rearrangement distance $d(T, A \oplus A \oplus \ldots \oplus A)$ is minimal. Part of this problem is to find an optimal labelling as $1, 2, \ldots$ or p of the p copies of each gene, so that all n copies labelled "1" are in one part of $A \oplus A \oplus \ldots \oplus A$ and all those labelled "2" are in a separate part, and so on. The genome A represents the ancestral genome at the moment immediately preceding the *polyploidization* event giving rise to $A \oplus A \oplus \ldots \oplus A$.

Warren also provided an efficient algorithm for the solution of genome aliquoting[50], though the complexity of this problem has not yet been established.

5 Kinds of Data

The analyses in the preceding subsections require complete and accurate linear gene orders along each chromosome. Experimental research, however, may sometimes only result in a partial ordering of the genes. Another problem is due to incompletely assembled genomes, where the linear order on a chromosome may be fragmented into many smaller linear orders called contigs. Finally, gene order data may contain errors, so that a proportion of genes are in the wrong place on a chromosome or even on the wrong chromosome. In this section we describe how the requirements of rearrangement theory may be reconciled with these kinds of data by using the rearrangement analysis itself to upgrade the imperfect data. Rather than reconstructing ancestral gene orders, then, we are reconstructing aspects of the structure of present-day ones.

5.1 Partially Ordered Genomes

The representation of a genome as a set of totally ordered chromosomes must often be weakened in the case of real data, where experimental data only suffice to partially order the set of genes on a chromosome. Maps of genes or other markers produced by recombination analysis, physical imaging and other methods, no matter how highly resolved, inevitably are missing some (and usually most) genes or markers and fail to order some pairs of neighbouring genes with respect to each other. Even at the ultimate level of resolution, that of genome sequences, the application of different gene-finding protocols usually gives maps with different gene content.

The biological concepts and usual computational methods of genome rearrangement, however, pertain only to totally ordered sets of genes and are meaningless in the context of partial orders. Our approach is to extend genome rearrangement theory to the more general context where gene order knowledge is represented by partial orders rather than total orders, i.e., directed acyclic graphs (DAGs) instead of linear graphs[51-52]. The use of DAGs reflects uncertainty of the gene order on chromosomes in the genomes of many advanced organisms. This may be due to lack of resolution, where several genes are mapped to the same chromosomal position, to missing data from some of the datasets used to compile a gene order, and/or to conflicts between these datasets.

We construct the chromosomal DAGs for each species from two or more incomplete datasets, or from a single low-resolution dataset. The frequent lack of order information in each dataset, due to missing genes or missing order information, is converted into parallel subpaths within each chromosomal DAG in a straightforward manner.

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. This is illustrated in Fig.3.

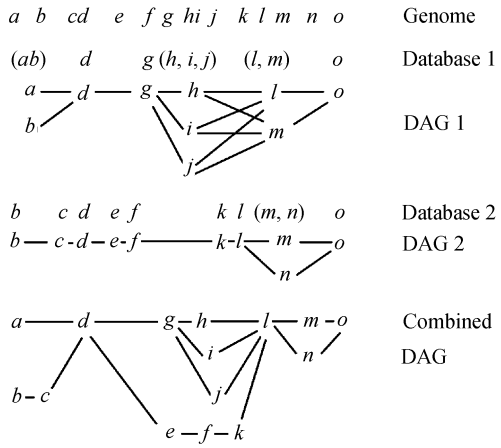


Fig.3. Construction of DAGs from individual databases each containing partial information on genome, due to missing genes and missing order information, followed by construction of combined DAG representing all known information on the genome. All edges directed from left to right.

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 datasets (as in the datasets on our simulated genomes), 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 minimal DAG whose vertex set is the union of all gene sets on that chromosome in the contributing datasets, 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 in Fig.3.

The rearrangement problem is then *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 general 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 in Fig.4.

Comparison of DAGs for genome comparison are generally hard problems[53], but considerable work has been done on heuristics[51-52], approximations[54] and generalizations[53].

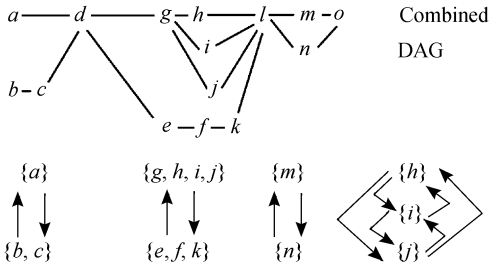


Fig.4. Edges added to DAG to obtain DG containing all linearization as paths (though not all paths in the DG are linearizations of the DAG!). Each arrow represents a set of directed edges, one from each element in one set to each element of the other set.

5.2 Fragmented Gene Orders

The sequencing coverage of many genomes is not sufficient to produce completely assembled genomes. Instead the published and archived data remain in *contig* form, i.e., continuous sequence of (usually much) smaller length than a chromosome. The price paid for increasing phylogenetic coverage in genome sequencing is the decreasing sequence coverage for each genome.

While such data may be adequate for many types of comparative genomic studies, they are not directly usable as input to genome rearrangement algorithms, since these algorithms require whole genome data, i.e., complete representations of each chromosome in terms of gene order, conserved segment order, or some other marker order, in order to calculate the rearrangement distance d between two genomes. Items whose chromosomal location is unknown cannot be part of the input.

Is there any way to use genome rearrangement algorithms to compare genomes available in contig form only[49,55-56]? Our suggestion[30] is to use the contigs directly in the rearrangement algorithms as if they were chromosomes. This biases the distance because it now counts extra fusion/fission operations necessary to compare genomes with different numbers of chromosomes and fragments of chromosomes. This bias must be corrected if, for example, we wish to use the results in a

distance matrix to input to a phylogenetic analysis.

Suppose we have one completely assembled genome B and another, A, in contig form only. The basic idea is that if we treat each contig as a chromosome, a rearrangement algorithm will automatically carry out a number of "fusions" to assemble the χ_A contigs in A into a small number of inferred chromosomes equal to the number χ_B in B, in calculating d.

Thus, when we use a rearrangement algorithm to compare a genome A in contig form with an assembled genome B, obtaining a preliminary distance d', it may seem appropriate to correct this to

$$d = d' - (\chi_A - \chi_B). \tag{8}$$

However, we cannot simply substitute correction (8) into (4) or (6) to estimate τ. Even if the number C of contig fusions is fixed, we know that these fusions are done in such a way as to minimize d'. To take account of this, we should only remove a proportion α of $C = \chi_A - \chi_B$ from d'. The natural hypothesis is that the effect of the C fragmentation operations creating C extra contigs would have the same effect as the addition of C of any other types of rearrangement to the genome, namely increasing d by approximately $\frac{\mathrm{d\,E}(d)}{\mathrm{d}\tau}C$.

To verify this, Muñoz[30] undertook a series of simulations, starting from an initial genome B containing 8867 genes (to mimic empirical data from *Drosophila*) in $\chi_B = 6$ chromosomes, generating 100 rearranged genomes, each through τ random rearrangements applied to B to produce a new genome, and each then fragmented into χ_A contigs. The average results for d' are summarized in Fig.5(a). First the linearity of the response to increasing χ_A is clear, indicating that (8) should be replaced by

$$d = d' - \alpha(\tau)(\chi_A - \chi_B), \tag{9}$$

where $\alpha(\tau)$ is a decreasing function of the number of rearrangements τ. As can be seen from Fig.5(b) this decrease only approximately parallels the theoretical derivative of d.

Given d', then, we can solve (5) and (9) simultaneously to find τ and d, since $n, \lambda_1, \lambda_2, \chi_A$ and χ_B are known, as is the dependence of α on τ.

A similar analysis can be carried out when both genomes are given in contig form.

5.3 Noisy Genomes

With some methodologies, the construction of gene orders is very vulnerable to errors. A typical problem involves ambiguous homology or paralogy, due to WGD and other duplication processes, leading to the risk of

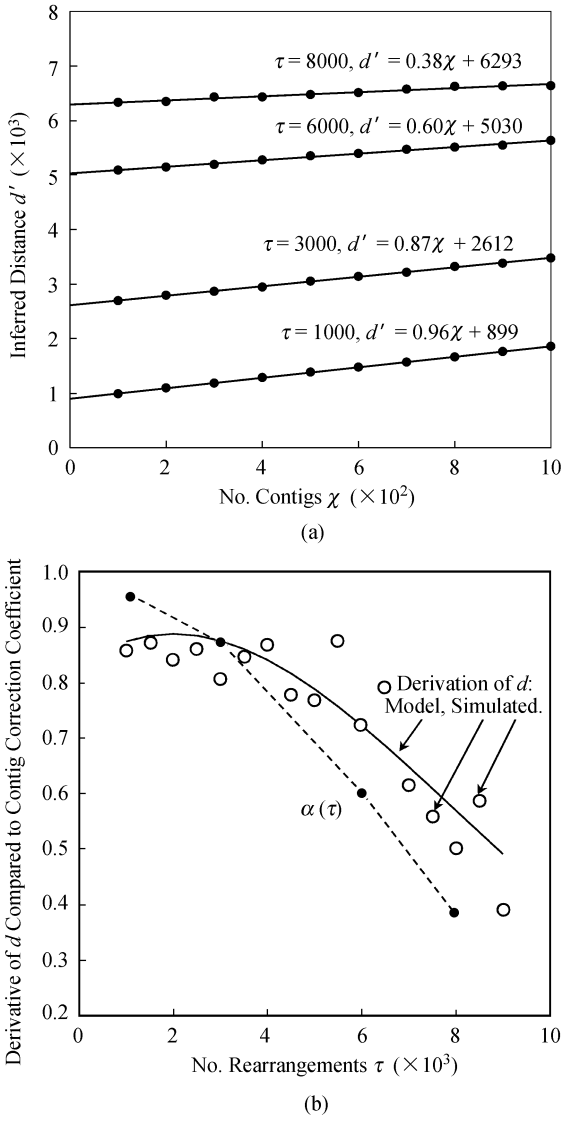


Fig.5. For genomes generated by $\tau = 1000, 3000, 6000$ or 8000 rearrangements, broken into $\chi = 100, 200, \ldots, 1000$ contigs: (a) the relationship between uncorrected genomic distance d' and χ and equations of trend lines. (b) The parameter α as a function of τ (filled dots and dotted line). Solid line represents the derivative of $\mathrm{E}(d)$ in (5), while the open dots represent the simulated values of the derivative, calculated from the data presented in Fig.1.

matching up inappropriate pairs of genes as orthologs in the two genomes[22]. These problems tend to artifactually disrupt long runs of consecutive genes in both genomes, and increase the number of shorter runs, often consisting of only one or two genes.

When many rearrangements have intervened since the common ancestor, it may be unclear whether any particular one of the increasing number of short runs is due to error or to rearrangement. These considerations suggest the principle that inferences that depend on the position of a single gene should not be given as much

weight as inferences that are supported by longer runs of genes.

In [57-58], we proposed the following strategy: First, construct a set of *pre-strips*, which are certain short common subsequences of one chromosome from each genome; second, extract from this set a subset of mutually compatible (non-intersecting) pre-strips containing a maximum number of genes; third, add to this subset any genes that do not increase the rearrangement distance between the gene orders; fourth, assemble the runs of genes.

First we define *strips, pre-strips* and *pure strips*. Consider any $l \geqslant 2$ consecutive contiguous genes on a chromosome in one genome. If the same l genes are consecutive on a chromosome in the other genome, with the same (or reverse) order and with each gene having the same (or opposite) orientation in both genomes, they constitute a *forward strip* (*reverse strip*) of *length* l.

We will search for *pre-strips* in the two genomes, relying on the subsequent analyses to eliminate the disrupting genes and thus reveal the "underlying" strips. This is illustrated in Fig.6. A pre-strip P is a common subsequence, or a reverse common subsequence, of the genes on the two chromosomes, such that there is no other gene of appropriate orientation on both chromosomes that is between two successive genes in P. For example, if AB is a pre-strip, then there does not exist C such that ACB is a pre-strip. A pre-strip satisfies the same definition as a strip, except that the genes need not be contiguous. A pre-strip that is a strip in the original genome data, and is not contained in another strip, is called a *pure strip*.

Remark. *Strips* are defined relative to the current state of the two genomes, either before, during or after reducing their size, but *pre-strips* and *pure strips* are defined in terms of the original genome data only.

ORIGINAL		**REDUCED**	
Genome 1	**Genome 2**	**Genome 1**	**Genome 2**
abcdef	lbcdpz	abcd	lbcdz
lmnoprq	-x-q-o-m	lmoq	-q-o-m
wxyz	we-fry	wyz	wy
	na		a
Pre-strips	**Pure strip**	**Strips**	**Singletons not**
bcd, bc, cd,	bcd	bcd, moq, wy	**in pre-strips**
moq, mo, oq,			**but compatible**
wy, lp			a, l, z
Common subsequences		**Discarded as noise**	
not pre-strips		e,f,n,p,r,x	
bd, mq			

Fig.6. Strips and pre-strips. "-" indicates different orientation markers in two genomes.

Maximal Strip Recovery (MSR) Problem: *Given two genomes as described above, discard some subset of the genes, leaving only genes in disjoint strips* $S_1, \ldots, S_r$ *of lengths* $w_1, \ldots, w_r$*, respectively, in the genomes thus reduced, such that* $\sum_{i=1}^{r} w_i$ *is maximized.*

We say two pre-strips P and Q are *in conflict* if they share at least one gene or if one pre-strip, say P, contains a gene between two successive genes, in either genome, in the other pre-strip, Q. Otherwise P and Q are *compatible*.

The MSR problem corresponds to our previously stated goal of constructing a set of compatible strips containing as much of the data as possible.

Every pre-strip P has a unique representation as a string of p's and 1's, where a p represents a pure strip and a 1, called a *singleton*, represents a marker not in a pure strip. Moreover,

Proposition 5.1. *Any pre-strip can be uniquely represented by a sequence of terms of form* p, 11, 1p, p1, 111 *and* 1p1.

Proposition 5.2. *All possible strips that can be formed by the deletion of genes from two genomes, and that can be part of a solution to the MSR problem, are pre-strips of these genomes.*

Consequently, it suffices to consider only pre-strips of the forms mentioned in Proposition 5.1. All such pre-strips can be calculated by an algorithm requiring $O(n^4)$ time in the worst case. In practice, the running time is far less.

Once we have these two propositions, one way to find a solution to the MSR problem is to create a compatibility graph with the pre-strips as vertices, weighted by the number of genes in them, and edges between compatible pre-strips, followed by the application of a Maximum Weight Clique (MWC) algorithm. We implemented recent versions of MWC[59-60] and found we could solve realistic versions of the MSR for several hundred genes and pre-strips only if we restricted the elements of pre-strips to be relatively close on the chromosome[57].

Another approach to MSR is via the conflict graph, which is simply the complement of the compatibility graph, and a maximum weight independent set (MWIS) problem. We found that by taking advantage of the source of the incompatibilities in the chromosome-based data, we could effect a natural decomposition of the graph into connected components that are mostly interval graphs, or slight distortion of interval graphs. Because efficient algorithms are available for MWIS on interval graphs, this allows us to solve relatively large instances of the problem extremely efficiently[58].

A number of papers have recently proposed generalizations of MSR and analyzed its complexity[61-64]. It would be of great interest if this approach were integrated with sequence level methods of identifying orthologous segments or anchor regions[17-18].

6 Generalized Gene Adjacency

Given two highly diverged gene orders, it may be difficult to decide if some set of genes are close enough in both genomes to infer some ancestral proximity or some functional relationship. There are a number of formal criteria for gene clustering in two or more organisms, giving rise to cluster detection algorithms and statistical tests for the significance of clusters. These methods, comprehensively reviewed by Hoberman and Durand[65], all depend on arbitrary parameters that control, in different ways, the number of genes and the proximity of these genes on the chromosome in order to be considered a cluster.

Most clustering criteria, however, do not take account of gene order *within the cluster*, but if the genes in a cluster are in the same order in both genomes, they suggest a closer relationship than if one order appears random with respect to the other. Few attempts have been made to take order into account. In [66] we made the following definitions: Suppose we have identified k genes that form a cluster in both genome A and genome B. Number the clustered genes in genome A in order from 1 to k (ignoring any intervening genes that are not in the scope of the cluster in genome B) and let $g_1, \ldots, g_k$ be the order of these same genes in genome B. Similarly, re-number the genes from 1 to k according to their order on genome B, and let $h_1, \ldots, h_k$ be the order of these same genes in genome A.

1) The breakpoint metric (BAD):

$$BAD = \#_{(i=1,\ldots,k-1)}\{|g_i - g_{i+1}| > 1\}$$
$$= \#_{(i=1,\ldots,k-1)}\{|h_i - h_{i+1}| > 1\},$$

the number of times a pair of genes adjacent in the cluster in one genome is not adjacent in the other. Were genes $1, \ldots, k$ the only genes in the genomes, then BAD would just be the unsigned breakpoint distance.

2) The maximum adjacency disruption criterion (MAD):

$$MAD = \max_{i=1,\ldots,k-1}\{\max\{|g_i - g_{i+1}|, |h_i - h_{i+1}|\}\},$$

the maximum, over all pairs of adjacent genes in the cluster in either genome, of the difference in their positions in the gene order in the cluster in the other genome. A low value of MAD means that no gene in the cluster has drifted far from its position in the ancestral genome.

3) The summed adjacency disruption criterion (SAD):

$$SAD = \sum_{i=1,\ldots,k-1}\{|g_i - g_{i+1}| + |h_i - h_{i+1}|\},$$

the sum, over all pairs of adjacent genes in the cluster in both genomes, of the difference in their positions in the gene order in the cluster in the other genome. This measures the overall movement of genes within the cluster from their positions in the ancestral genome.

Values for MAD, for $k \leqslant 13$ are known and can be used for statistical testing. For large k, modelling suggests that

$$\Pr(MAD) \leqslant a \approx \beta(2-\beta)^{2k}, \tag{10}$$

where $\beta = a/k$.

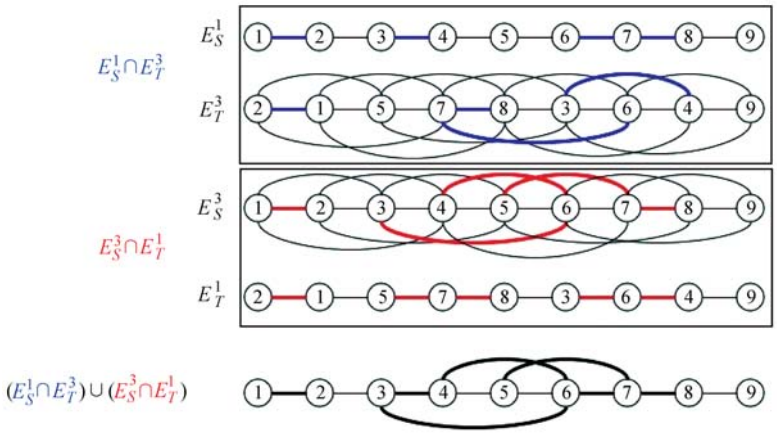


Fig.7. Determination of $(1, 3)$ clusters (or (3,1) clusters).

More recently we have introduced the notion of *generalized gene adjacency*[32,67-68]. We say two genes are (i,j)-adjacent if they are separated by $i-1$ genes on a chromosome in either one of the genomes and $j-1$ genes in the other. We define a (θ,ψ) cluster in terms of a graph where the genes are vertices and edges are drawn between those (i,j)-adjacent gene pairs where $\min(i,j) < \min(\theta,\psi)$ and $\max(i,j) < \max(\theta,\psi)$. Then the connected components of the graph are the (θ,ψ) clusters, as illustrated in Fig.7. Generalized adjacency clusters embody gene order considerations within the cluster in that they cannot have two genes close together in one genome but far apart in the other, although the cluster could be very large.

What value should we assign θ (and ψ)? To answer this, we first broaden the problem by defining a wide class of similarities (or equivalently, distances) between two genomes in terms of weights on the (i,j)-adjacencies, namely any system of fixed-sum, symmetric, non-negative weights ω non-increasing in i and j. This is the most general way of representing decreasing weight with increasing separation of the genes on the chromosome. Thus, given two genomes S and T with the same genes, let ω_{ij} be the **weight** on two genes that are (i,j)-adjacent, such that

1) $0 \leqslant \omega_{ij} = \omega_{ji}$, $i,j \in \{1,2,\ldots,n-1\}$,

2) $\sum_{i=1}^{n-1}\sum_{j=1}^{n-1}\omega_{ij} = 1$,

3) $\omega_{i,j} \geqslant \omega_{k,l}$ if

   (a) $\max(i,j) < \max(k,l)$, or

   (b) $\max(i,j) = \max(k,l)$ and $\min(i,j) < \min(k,l)$.

We define the **distance** between two genomes S and T as

$$d(S,T) = 2n - 2 - \sum_{i=1}^{n-1}\Big(n_{ii}\omega_{ii} + \sum_{j=1}^{n-1} n_{ij}\omega_{ij}\Big), \quad (11)$$

where n_{ij} is the total number of gene pairs (x,y) that are i-adjacent in S and j-adjacent in T. In any pair of genomes, we then wish to maximize the sum of the weights, which essentially maximizes the sensitivity of the criterion. Our main result is a theorem showing that the solution reduces to a uniform weight on gene separations up to a certain value of both θ and ψ, and zero weight on larger separations.

Theorem 6.1. *Let* $\alpha_k = \lfloor \frac{\sqrt{1+8(k-1)}+1}{2} \rfloor$. *The weight* ω *that minimizes* $d(S,T)$ *has*

$$\omega_{ij} = \begin{cases} \begin{cases} \frac{1}{k^*}, & \text{if } i < \alpha_{k^*},\ j \leqslant i, \text{or} \\ & i = \alpha_{k^*},\ j \leqslant k^* - \frac{i(i-1)}{2}, \end{cases} \\ 0, & \text{otherwise,} \end{cases} \quad (12)$$

where k^* *is a natural number and maximizes the function*

$$f(k) = \frac{1}{k}\Big[\sum_{i=1}^{\alpha_k - 1}\sum_{j=1}^{i}(n_{ij}+n_{ji}) + \sum_{j=1}^{k-\frac{1}{2}\alpha_k(\alpha_k-1)}(n_{\alpha_k j} + n_{j\alpha_k})\Big], \quad (13)$$

where n_{ij} *is the number of gene pairs* i*-adjacent on* S *and* j*-adjacent on* T. (*See Fig.*8 *for the* 2*-dimensional area measured by* k^*.)

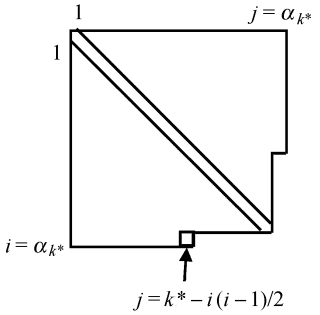


Fig.8. k is augmented from left to right, starting at the top row, in the lower triangle including the diagonal. Values of ω_{ij} in the upper triangle determined by symmetry.

We can set $\theta = \psi = \lfloor \frac{\sqrt{1+8(k^*-1)}+1}{2} \rfloor \approx \sqrt{2k^*}$ and use $\mathrm{E}[k^*]$, as function of n, to find the natural value for the cut-off parameters $\theta = \psi$ in the uniform weight-based distance.

We can use the fact that the n_{ij} are Poisson with parameter $\mathrm{E}(n_{ij}) = \frac{2(n-i)(n-j)}{n(n-1)}$ to show:

Theorem 6.2. *Let* n_{ij} *be the number of gene pairs* i*-adjacent on* S *and* j*-adjacent on* T, *then the* $f(k)$ *in Theorem* 6.1 *satisfies*

$$\mathrm{E}[f(k)] \to \Big(2 - \frac{\alpha}{n}\Big)^2$$
$$\mathrm{Var}[f(k)] \to \frac{8}{\alpha^2}\Big(1 + \frac{2}{\alpha} - \frac{2}{\alpha^2}\Big) \quad (14)$$

as $n \to \infty$, *where* $\alpha = \lfloor \frac{\sqrt{1+8(k-1)}+1}{2} \rfloor$.

Though we cannot find k^* analytically, we can characterize it quite well using simulations[68].

A second set of "natural" parameter values to serve as an upper bound on the meaningful choices of θ and ψ are the percolation thresholds of the (θ, ψ) clusters. Beyond these values, tests of significance are not meaningful because all clusters rapidly coalesce together. It is no longer significant to find very large gene clusters.

Percolation has been studied for max-gap clusters[69], but the main analytical results on percolation pertain to completely random (Erdös-Rényi) graphs. The graphs associated with (θ, ψ) clusters manifest delayed percolation, so the use of Erdös-Rényi percolation values would be a "safe" but conservative way of avoiding dangerously high values of the parameters.

It was established by Erdös and Rényi[70-72] that for random graphs where edges are independently present between pairs of the n vertices with probability p, the percolation threshold is $p = 1/n$.

We note that the percolation of the generalized adjacency graph is delayed considerably compared to unconstrained Erdös-Rényi graphs with the same number of edges, as may be seen in Fig.9. To understand what aspect of the generalized adjacency graphs is responsible for this delay, we also simulated random graphs of bandwidth $\leqslant \theta$, since this constraint is an important property of generalized adjacency. It can be seen in Fig.9, that the limited bandwidth graphs also show delayed percolation, but less than half that of generalized adjacency graphs.

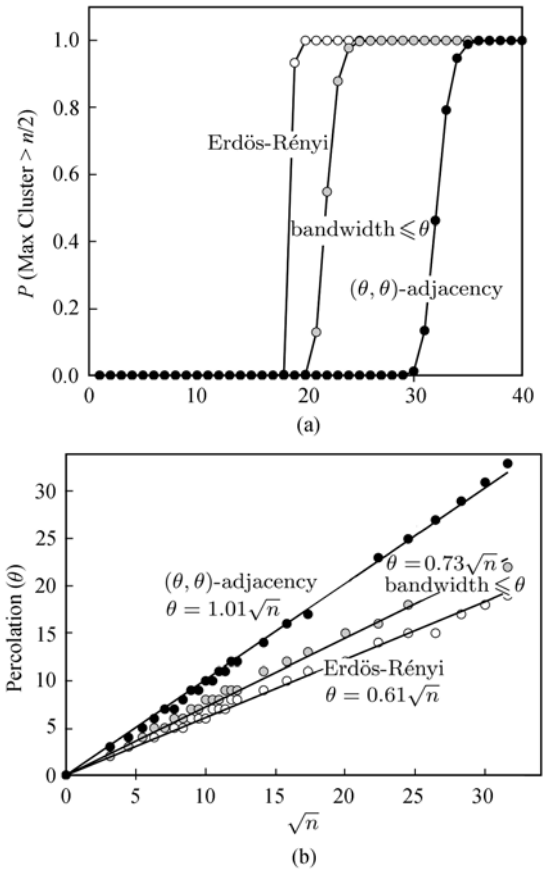


Fig.9. (a) Simulation with genome length $n = 1000$, with $2\theta^2$ edges in each graph, showing delayed percolation of generalized adjacency graphs with respect to Erdös-Rényi graphs. Bandwidth-limited graphs are also delayed but much less so. (b) Percolation point as a function of $\sqrt{n}$, again with $2\theta^2$ edges per graph. Delay measured by coefficient of $\sqrt{n}$ in equation for trend line.

As a control on our simulations, it is known (cf. [73]) that Erdös-Rényi graphs with rn edges, with r somewhat larger than $1/2$ have a cluster of size $(4r - 2)n$. Our percolation criterion is that one cluster must have at least $n/2$ vertices. Solving this, we get $r = 0.625$. This means that the $2\theta^2$ edges we use in each of our simulated graphs must be the same as $0.625n$, suggesting that $\theta = 0.56\sqrt{n}$, compared to the $0.61\sqrt{n}$ we found in our limited simulations.

The two major mathematical and computational problems arising from this work are the analytical prediction of the random variable k^*, and the accounting for the delay in percolation behaviour of generalized adjacency graphs beyond the bandwidth effect.

7 Conclusions

The reconstruction of ancestral genomes involves a diverse set of algorithmic, graph theoretical, simulation modelling, statistical and probability approaches, many of which have been extensively investigated, others less so, but all of which are suggestive of further investigation and application. Although we have not dwelt on this in this chapter, the developments in this field have been continually provoked by and refined by applications to gene order data, from organelles like mitochondria and chloroplasts, through prokaryotic organisms like bacteria, to protists, yeasts, algae and higher plants, insects and higher animals, including vertebrates at all levels and especially mammals.

References

[1] Venter J C, Adams M D, Myers E W, Li P W, Mural R J, Sutton G G *et al.* The sequence of the human genome. *Science*, 2001, 291(5507): 1304-1351.

[2] Mouse Genome Sequencing Consortium. Initial sequencing and comparative analysis of the mouse genome. *Nature*, 2002, 420(6915): 520-562.

[3] Sankoff D, Leduc G, Antoine N, Paquin B, Lang B F, Cedergren R. Gene order comparisons for phylogenetic inference: Evolution of the mitochondrial genome. *Proc. the National Academy of Sciences USA*, 1992, 89(14): 6575-6579.

[4] Sankoff D. Edit distance for genome comparison based on non-local operations. In *Proc. the Third Annual Symposium on Combinatorial Pattern Matching* (*CPM 1992*), Tucson, USA, April 29-May 1, 1992, pp.121-135.

[5] Cosner M E, Jansen R K, Moret B M E, Raubeson L A, Wang L-S, Warnow T, Wyman S. An Empirical Comparison

of Phylogenetic Methods on Chloroplast Gene Order Data in Campanulaceae. Comparative Genomics: Empirical and Analytical Approaches to Gene Order Dynamics, Map Alignment, and the Evolution of Gene Families, Sankoff D, Nadeau J (eds.), Dordrecht: Kluwer Academic Publishers, 2000, pp.99-121.

[6] Ajana Y, Lefebvre J-F, Tillier E R M, El-Mabrouk N. Exploring the set of all minimal sequences of reversals — An application to test the replication-directed reversal hypothesis. In *Proc. the Second International Workshop on Algorithms in Bioinformatics* (*WABI 2002*), Rome, Italy, Sept. 17-21, 2002, pp.300-315.

[7] Hannenhalli S, Pevzner P A. Transforming cabbage into turnip: Polynomial algorithm for sorting signed permutations by reversals. *Journal of the ACM*, 1999, 46(1): 1-27.

[8] Caprara A. Sorting permutations by reversals and Eulerian cycle decompositions. *SIAM Journal on Discrete Mathematics*, 1999, 12(1): 91-110.

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

[10] Sankoff D. Mechanisms of genome evolution: Models and inference. *Bulletin of the International Statistical Institute*, 1989, 47(3): 461-475.

[11] Sturtevant A H, Novitski E. The homologies of the chromosome elements in the genus *Drosophila*. *Genetics*, 1941, 26(5): 517-541.

[12] Hannenhalli S, Pevzner P A. Transforming men into mice (polynomial algorithm for genomic distance problem). In *Proc. the Thirty-Sixth Annual Symposium on Foundations of Computer Science* (*FOCS 1995*), Milwaukee, USA, Oct. 23-25, 1995, pp.581-592.

[13] Tesler G. Efficient algorithms for multichromosomal genome rearrangements. *Journal of Computer and System Sciences*, 2002, 65(3): 587-609.

[14] Sankoff D, Blanchette M. Multiple genome rearrangement and breakpoint phylogeny. *Journal of Computational Biology*, 1998, 5(3): 555-570.

[15] El-Mabrouk N, Bryant D, Sankoff D. Reconstructing the pre-doubling genome. In *Proc. the Third Annual International Conference on Computational Molecular Biology* (*RECOMB*), Lyon, France, April 11-14, 1999, pp.154-163.

[16] El-Mabrouk N, Sankoff D. The reconstruction of doubled genomes. *SIAM Journal on Computing*, 2003, 32(2): 754-792.

[17] Pevzner P, Tesler G. Human and mouse genomic sequences reveal extensive breakpoint reuse in mammalian evolution. *Proceedings of the National Academy of Sciences USA*, 2003, 100(13): 7672-7677.

[18] Kent W J, Baertsch R, Hinrichs A, Miller W, Haussler D. Evolution's cauldron: Duplication, deletion, and rearrangement in the mouse and human genomes. *Proceedings of the National Academy of Sciences USA*, 2003, 100(20): 11484-11489.

[19] Kent W J, Sugnet C W, Furey T S, Roskin K M, Pringle T H, Zahler A M, Haussler A D. The Human genome browser at UCSC. *Genome Research*, 2002, 12(6): 996-1006.

[20] Mazowita M, Haque L, Sankoff D. Stability of rearrangement measures in the comparison of genome sequences. *Journal of Computational Biology*, 2006, 13(2): 554-566.

[21] Sinha A U, Meller J. Sensitivity analysis for reversal distance and breakpoint reuse in genome rearrangements. *Pacific Symposium on Biocomputing*, 2008, 13: 37-38.

[22] Jiang T. Some algorithmic challenges in genome-wide ortholog assignment. *J. Comput. Sci. & Technol.*, 2010, 25(1): 42-52.

[23] Tannier E, Zheng C, Sankoff D. Multichromosomal median and halving problems under different genomic distances. *BMC Bioinformatics*, 2009, 10: 120.

[24] Fertin G, Labarre A, Rusu I, Tannier E, Vialette S. Combinatorics of Genome Rearrangements. Cambridge, Massachusetts: The MIT Press, 2009.

[25] Yancopoulos S, Attie O, Friedberg R. Efficient sorting of genomic permutations by translocation, inversion and block interchange. *Bioinformatics*, 2005, 21(16): 3340-3346,

[26] Bergeron A, Mixtacki J, Stoye J. A unifying view of genome rearrangements. In *Proc. the Sixth International Workshop on Algorithms in Bioinformatics* (*WABI 2000*), Zurich, Switzerland, Sept. 11-13, 2006, pp.163-173.

[27] Dalevi D, Eriksen N. Expected gene-order distances and model selection in bacteria. *Bioinformatics*, 2008, 24(11): 1332-1338.

[28] Eriksen N, Hultman A. Estimating the expected reversal distance after a fixed number of reversals. *Advances in Applied Mathematics*, 2004, 32(3): 439-453.

[29] Wang L-S, Warnow T. Distance-Based Genome Rearrangement Phylogeny. *J. Mol. Wvol.*, 2006, 63(4): 473-483.

[30] Muñoz A, Sankoff D. Rearrangement phylogeny of genomes in contig form. In *Proc. the Fifth International Symposium on Bioinformatics Research and Applications* (*ISBRA 2009*), Fort Lauderdale, USA, May 13-16, 2009, pp.160-172.

[31] Adam Z, Turmel M, Lemieux C, Sankoff D. Common intervals and symmetric difference in a model-free phylogenomics, with an application to streptophyte evolution. *Journal of Computational Biology*, 2007, 14(4): 436-445.

[32] Zhu Q, Adam Z, Choi V, Sankoff D. Generalized gene adjacencies, graph bandwidth, and clusters in yeast evolution. *Transactions on Computational Biology and Bioinformatics*, 2009, 6(2): 213-220.

[33] Tannier E. Yeast ancestral genome reconstructions: The possibilities of computational methods. In *Proc. the 7th Ann. RECOMB Satellite Workshop on Comparative Genomics* (*RECOMB CG 2009*), Budapest, Hungary, Sept. 27-29, 2009, pp.1-12.

[34] Sankoff D, Blanchette M. The median problem for breakpoints in comparative genomics. In *Proc. the Third Annual International Conference on Computing and Combinatorics* (*COCOON 1997*), Shanghai, China, Aug. 20-22, 1997, pp.251-263.

[35] Bader D, Moret B M. GRAPPA runs in record time. *HPCwire*. November 23, 2000, 9(47).

[36] Siepel A C. Exact algorithms for the reversal median problem [Master's Thesis]. University of New Mexico, 2001.

[37] Caprara A. On the practical solution of the reversal median problem. In *Proc. the First International Workshop on Algorithms in Bioinformatics* (*WABI 2001*), Aarhus, Denmark, Aug. 28-31, 2001, pp.238-251.

[38] Bourque G, Pevzner P A. Genome-scale evolution: Reconstructing gene orders in the ancestral species. *Genome Research*, 2002, 12(1): 26-36.

[39] Xu A W. A fast and exact algorithm for the median of three problem — A graph decomposition approach. In *Proc. Sixth Annual RECOMB Satellite Workshop Comparative Genomics* (*RECOMB CG 2008*), Paris, France, Oct. 13-15, 2008, pp.184-197.

[40] Xu A W, Sankoff D. Decompositions of multiple breakpoint graphs and rapid exact solutions to the median problem. In *Proc. the Eighth International Workshop on Algorithms in Bioinformatics* (*WABI 2008*), Karlsruhe, Germany, Sept. 15-17, 2008, pp.25-37.

[41] Adam Z, Sankoff D. A statistically fair comparison of ancestral genome reconstructions, based on breakpoint and rearrangement distances. In *Proc. the Seventh Annual RECOMB Satellite Workshop on Comparative Genomics* (*RECOMB CG 2009*), Budapest, Hungary, Sept. 16-18, 2009, pp.193-204.

[42] Adam Z, Sankoff D. The ABCs of MGR with DCJ. *Evolutionary Bioinformatics Online*, 2008, 4: 69-74.

[43] Warren R, Sankoff D. Genome halving with double cut and join. *Journal of Bioinformatics and Computational Biology*, 2009, 7(2): 357-371.

[44] Mixtacki J. Genome halving under DCJ revisited. In *Proc. the Fourteenth Annual Conference on Computing and Combinatorics* (*COCOON*), Dalian, China, June 27-29, 2008, pp.276-286.

[45] Blin G, Chauve C, Fertin G, Rizzi R, Vialette S. Comparing genomes with duplications: A computational complexity point of view. *Transactions on Computational Biology and Bioinformatics*, 2007, 4(4): 523-534.

[46] Zheng C, Zhu Q, Sankoff D. Genome halving with an outgroup. *Evolutionary Bioinformatics*, 2006, 2(13): 319-326.

[47] Sankoff D, Zheng C, Zhu Q. Polyploids, genome halving and phylogeny. *Bioinformatics*, 2007, 23(13): i433-i439.

[48] Zheng C, Zhu Q, Adam Z, Sankoff D. Guided genome halving: Hardness, heuristics and the history of the Hemiascomycetes. *Bioinformatics*, 2008, 24(13): i96-i104.

[49] Sankoff D, Zheng C, Wall P K, dePamphilis C, Leebens-Mack J, Albert V A. Towards improved reconstruction of ancestral gene order in angiosperm phylogeny. *Journal of Computational Biology*, 2009, 16(10): 1353-1367.

[50] Warren R, Sankoff D. Genome aliquoting with double cut and join. *BMC Bioinformatics*, 2009, 10: S2.

[51] Zheng C, Lenert A, Sankoff D. Reversal distance for partially ordered genomes. *Bioinformatics*, 2005, 21(Suppl. 1): i502-i508.

[52] Zheng C, Sankoff D. Genome rearrangements with partially ordered chromosomes. *Journal of Combinatorial Optimization*, 2006, 11(2): 133-144.

[53] Blin G, Blais E, Hermelin D, Guillon P, Blanchette M, El-Mabrouk N. Gene maps linearization using genomic rearrangement distances. *Journal of Computational Biology*, 2007, 14(4): 394-407.

[54] Chen X, Cui Y. An approximation algorithm for the minimum breakpoint linearization problem. *Transactions on Computational Biology and Bioinformatics*, 2009, 6(3): 401-409.

[55] Gaul E, Blanchette M. Ordering partially assembled genomes using gene arrangements. In *Proc. the Fourth Annual Workshop on Comparative Genomics* (*RECOMB CG 2006*), Montreal, Canada, Sept. 24-26, 2006, pp.113-128.

[56] Bhutkar A, Russo S, Smith T F, Gelbart W M. Techniques for multi-genome synteny analysis to overcome assembly limitations. *Genome Informatics*, 2006, 17(2): 152-161.

[57] Zheng C, Zhu Q, Sankoff D. Removing noise and ambiguities from comparative maps in rearrangement analysis. *Transactions on Computational Biology and Bioinformatics*, 2007, 4(4): 515-522.

[58] Choi V, Zheng C, Zhu Q, Sankoff D. Algorithms for the extraction of synteny blocks from comparative maps. In *Proc. the Seventh International Workshop on Algorithms in Bioinformatics* (*WABI 2007*), Philadelphia, USA, Sept. 8-9, 2007, pp.277-288.

[59] Ostergard P R J. A new algorithm for the maximum-weight clique problem. *Nordic Journal of Computing*, 2001, 8(4): 424-436.

[60] Kumlander D. A new exact algorithm for the maximum-weight clique problem based on a heuristic vertex-coloring and a backtrack search. In *The Fourth European Congress of Mathematics*, Stockholm, Sweden, June 27-July 2, 2004, MS. and Poster.

[61] Bulteau L, Fertin G, Rusu I. Maximal strip recovery problem with gaps: Hardness and approximation algorithms. In *Proc. the 20th Int. Symp. Algorithms and Computation* (*ISAAC 2009*), Hawaii, USA, Dec. 16-18, 2009, pp.710-719.

[62] Chen Z, Fu B, Jiang M, Zhu B. On recovering syntenic blocks from comparative maps. In *Proc. the Second Annual Int. Conf. Combinatorial Optimization and Applications* (*COCOA 2008*). St. John's, Canada, Aug. 21-24, 2008, pp.319-327.

[63] Jiang M. Inapproximability of maximal strip recovery. In *Proc. the 20th Int. Symp. Algorithms and Computation* (*ISAAC 2009*), Hawaii, USA, Dec. 16-18, 2009, pp.616-625.

[64] Wang L, Zhu B. On the tractability of maximal strip recovery. In *Proc. the Sixth Annual Conf. Theory and Applications of Models of Computation* (*TAMC 2009*), Changsha, China, May 18-22, 2009, pp.400-409.

[65] Hoberman R, Durand D. The incompatible desiderata of gene cluster properties. In *Proc. the Fifth Annual Workshop on Comparative Genomics* (*RECOMB CG 2005*), Dublin, Ireland, Sept. 18-20, 2005, pp.73-87.

[66] Sankoff D, Haque L. Power boosts for cluster tests. In *Proc. the Fifth Annual Workshop on Comparative Genomics* (*RECOMB CG*), Dublin, Ireland, Sept. 18-20, 2005. pp.121-130.

[67] Xu X, Sankoff D. Tests for gene clusters satisfying the generalized adjacency criterion. In *Proc. the Third Brazilian Symposium on Bioinformatics, Advances in Bioinformatics and Computational Biology* (*BSB 2008*), Santo Andre, Brazil, Aug. 28-30, 2008, pp.152-160.

[68] Yang Z, Sankoff D. Natural parameter values for generalized gene adjacency. In *Proc. the Seventh Annual RECOMB Satellite Workshop on Comparative Genomics* (*RECOMB CG*), San Diego, USA, Sept. 16-18, 2009, pp.13-23.

[69] Hoberman R, Sankoff D, Durand D. The statistical analysis of spatially clustered genes under the maximum gap criterion. *Journal of Computational Biology*, 2005, 12(8): 1083-1102.

[70] Erdös P, Rényi A. On random graphs. *Publicationes Mathematicae*, 1959, 6: 290-297.

[71] Erdös P, Rényi A. On the evolution of random graphs. *Publications of the Mathematical Institute of the Hungarian Academy of Sciences*, 1960, 5: 17-61.

[72] Erdös P, Rényi A. On the strength of connectedness of a random graph. *Acta Mathematica Scientia Hungary*, 1961, 12: 261-267.

[73] D'Souza R, Achlioptas D, Spencer J. Explosive percolation in random networks. *Science*, 2009, 323(5920): 1453-1455.

David Sankoff holds the Canada Research Chair in Mathematical Genomics in the Mathematics and Statistics Department at the University of Ottawa, and is cross-appointed to the Biology Department and the School of Information Technology and Engineering. His research interest is comparative genomics, particularly probability models, statistics and algorithms for genome rearrangements.

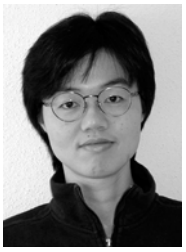
Chunfang Zheng studied biology at Beijing Sports University and computer science at the University of Ottawa, where she then obtained her Master's and Ph.D. degrees in biology. She is a postdoctoral fellow with Nadia El-Mabrouk at the Université de Montréal. She has worked on the comparison of partially ordered and noisy genomes and the incorporation of whole genome duplication descendants into gene order phylogeny.

Adriana Muñoz holds a Master's degree from the University of Alberta and is currently a computer science Ph.D. candidate interested in the comparison of incompletely assembled genomes.

Zhenyu Yang holds a Master's degree in computer science from Liaoning University in China and is currently a mathematics Ph.D. candidate interested in gene clusters in comparative genomics.

Zaky Adam holds a Master's degree from the University of Western Ontario and completed his computer science Ph.D. at the University of Ottawa on gene order phylogeny. He is currently a postdoctoral fellow with Kateryna Makova at Penn State University.

Robert Warren holds a Master's degree from the University of British Columbia and is currently a computer science Ph.D. candidate interested in algorithms for genome halving, genome aliquoting and related problems.



Vicky Choi did her undergraduate studies at the Chinese University of Hong Kong and her masters at the Hong Kong University of Science and Technology. She was awarded her Ph.D. degree in computer science from Rutgers University. She has been assistant professor at Virginia Tech since 2004, except for a sabbatical leave working at a quantum computing enterprise. Her main research areas are the design, implementation, and analysis of algorithms and quantum computing.

Qian Zhu did his undergraduate studies in the biochemistry program at the University of Ottawa. He is currently a graduate student in computer science at Princeton University.