Parts of the Problem of Polyploids in Rearrangement Phylogeny

Chunfang Zheng, Qian Zhu, and David Sankoff

Departments of Biology, Biochemistry, and Mathematics and Statistics,
University of Ottawa, Ottawa, Canada K1N 6N5
{czhen033,qzhu012,sankoff}@uottawa.ca

Abstract. Genome doubling simultaneously doubles all genetic markers. Genome rearrangement phylogenetics requires that all genomes analyzed have the same set of orthologs, so that it is not possible to include doubled and unduplicated genomes in the same phylogeny. A framework for solving this difficulty requires separating out various possible local configurations of doubled and unduplicated genomes in a given phylogeny, each of which requires a different strategy for integrating genomic distance, halving and rearrangement median algorithms. In this paper we focus on the two cases where doubling precedes a speciation event and where it occurs independently in both lineages initiated by a speciation event. We apply these to a new data set containing markers that are ancient duplicates in two yeast genomes.

1 Introduction

Basic rearrangement phylogeny methods require that the genomic content be the same in all the organisms being compared, so that every marker (whether gene, anchor, probe binding site or chromosomal segment) in one genome be identified with a single orthologous counterpart in each of the others, though adjustments can be made for a limited amount of marker deletion, insertion and duplication.

Many genomes have been shown to result from an ancestral doubling of the genome, so that every chromosome, and hence every marker, in the entire genome is duplicated simultaneously. Subsequently, the doubled genome evolves through mutation at the DNA sequence level and by chromosomal rearrangement, through intra- and interchromosomal movement of genetic material. This movement can scramble the order of markers, so that the chromosomal neighbourhood of a marker need bear no resemblance to that of its duplicate.

The present-day genome, which we refer to here as a doubling descendant, can be decomposed into a set of duplicate or near-duplicate markers dispersed among the chromosomes. There is no direct way of partitioning the markers into two sets according to which ones were together in the same half of the original doubled genome. Genomic distance or rearrangement phylogeny algorithms are not applicable to doubling descendants, since there is a two-to-one relationship between markers in the doubling descendant and related species whose divergence predates the doubling event, whereas these algorithms require a one-to-one correspondence.

G. Tesler and D. Durand (Eds.): RECOMB-CG 2007, LNBI 4751, pp. 162–176, 2007.

We have undertaken a program [11,9] of studying rearrangement phylogeny where doubling descendants are considered along with related unduplicated genomes. We believe there is no other computationally-oriented literature on this particular problem. To focus on the problem of marker ambiguity in doubling descendants, and to disentangle it from the difficulties of constructing phylogenies, we pose our computational problems only within the framework of the "small" phylogenetic problem, i.e., identifying the ancestral genomes for a given phylogeny that jointly minimize the sum of the rearrangement distances along its branches.

In Section 2, we outline a model for generating an arbitrary pattern of doubled descendants observed at the tips of a given phylogeny. Based on this model, we then present an simple algorithm for inferring the doubling status of the ancestral genomes in terms of an economical set of doubling events along the branches of the phylogeny. Once we have the ancestral doubling statuses, we can approach the actual rearrangement problem.

First, in Section 3, we identify three kinds of component of this problem for which algorithms already exist, one a calculation of the genomic distance between two given genomes with clearly identified orthologs, i.e., the minimum number of rearrangements necessary to transform one genome into another; the second a "halving" algorithm for inferring the genome of a doubled genome based on internal evidence from its modern descendant only, and the third a "medianizing" process for inferring an ancestral genome from its three neighbouring genomes in a binary branching tree.

In Section 4, we discuss our recent papers [11,9] on incorporating algorithms for the three components into an overall procedure for inferring ancestral genomes in the case of one doubling descendant and two related unduplicated genomes. The contribution of the present paper starts in Section 5 where we analyze two ways of relating genomes from two doubling descendants, one where they result from a single genome doubling event followed by a speciation, and the other where speciation precedes two genome doublings, one in each lineage. In Section 7, we apply these two methods to a large data set on yeast.

1.1 Terminology and Scope

In biology, the concept of genome doubling is usually expressed as tetraploidization or autotetraploidization, and the both the doubled genome and its doubling descendant are called tetraploid, even though, generally, the descendants soon undergo a process called (re-)diploidization and function as normal diploids, still carrying a full complement of duplicate markers that evolve independently of each other. Though unambiguous in biological context, implicit in this terminology are many assumptions that are not pertinent to our study. In the yeast data we study here, for example, *Saccharomyces cerevisiae* exists during most of its life cycle as a haploid, only sometimes as a diploid, while *Candida glabrata* exists uniquely as a haploid.

In our considerations, the key aspect of genome doubling is the global duplication of chromosomes and markers at the moment of doubling. Ploidy is not relevant in that in any organism that reproduces by meiosis or even by mitosis, the order of the markers on any of the haploid components (e.g., maternal versus paternal chromosomes) is essentially identical. There may be different alleles, or other local differences, but the order is basically invariant. Ongoing variation and evolution at the level of chromosomal structure in an individual or species are considered negligible in comparison with the major rearrangements that exist between genomes separated on an evolutionary time scale.

Although this paper is about polyploidy, then, we will rely largely on terminology independent of ploidy: genome doubling, doubling descendant, unduplicated genomes, genome halving.

The marker complement of a genome may also double by another process, allotetraploidization, or fusion of two different genomes, a kind of hybridization that is probably at least as important biologically as the doubling of a single genome we focus on in this paper. We do not consider this process here, for three reasons. One is our interest in exploring the essential difficulty in the mathematics of doubling, namely the complete ambiguity as to which set of duplicate markers were together in each of the two copies of the original genome. For hybrid doubled genomes, DNA sequence evidence from related but unduplicated genomes can generally resolve this ambiguity [5]. Second, hybrids require reticulate phylogenies which, though of interest themselves, constitute an unwanted layer of difficulty that we wish to keep separate at this stage. Finally, some of the most interesting doubling events (outside the plant kingdom), such as the ones hypothesized in the "2R" model of early vertebrate evolution or the well-established doubling in the ancestor of budding yeasts *Saccharomyces cerevisiae* and *Candida glabrata*, which furnish the empirical example for this paper, are usually treated as doubling of a single genome.

2 Generation and Inference of Polyploidy

Our algorithms require genomic sequence data or other high resolution marker data spanning the entire genome. This, of course, is only available in a limited number of phylogenetic domains within the eukaryotes, and then only from selected organisms. Our analysis may also benefit from information on doubling status not only about the sequenced or mapped genomes, but also from closely related organisms. Fortunately such information is much easier to obtain experimentally and to come by in the literature, though ancestral events often require inferential leaps based on the number of chromosomes or the distribution of the number of copies of each marker.

Our first task, given some mixture of doubling descendants and unduplicated genomes related by a phylogenetic tree, is to infer the doubling status of the all the ancestral genomes. Under the simplifying assumptions that all ploidies are powers of two and can only remain unchanged or change by a factor of two

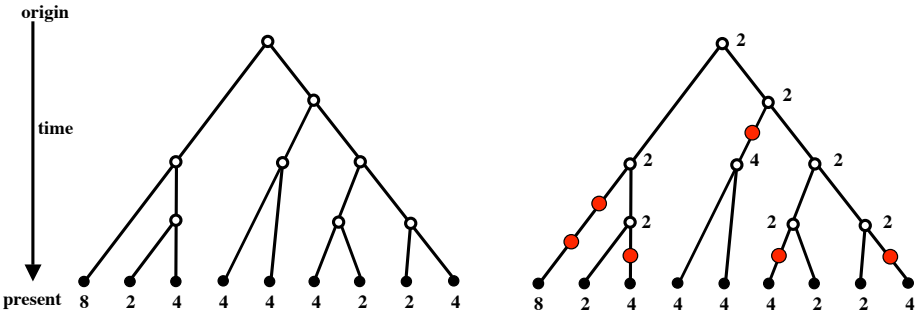

Fig. 1. Example of doubling inference problem. Genomes observed only for leaves (filled dots) of phylogeny. 2 = diploid unduplicated genome. Inferred doubling events indicated by red dots.

at each step, and the parsimony criterion that the number of doubling steps is to be minimized, the task is achieved by the recurrence

$$\Pi(v) = \min_{\text{daughter species } u \text{ of v}} \Pi(u)$$

at each ancestral vertex v of a phylogenetic tree, as depicted in Fig. 1.

Once Π is inferred, the doubling events may be inferred to occur on those branches of the tree where the Π differs at the two ends. This is also depicted in Fig.1. In the ensuing sections, we will illustrate the local configurations giving rise to various inference problems by highlighting appropriate portions of the tree in Fig. 1.

3 Existing Resources

Once we have inferred the doubling status of the ancestral genomes, how are we to approach our original problem: to reconstruct the marker order of the ancestral genomes and thus infer the cost of the phylogeny in terms of rearrangement events? Here we discuss some basic elements of the solution.

Genomic distance. Distance based on genomic structure $d(X, Y)$ is calculated by linear-time rearrangement algorithms for finding the minimum number of operations necessary to convert one genome X into another Y. Genomic distance is defined only between genomes of the same ploidy, as highlighted in the leftmost example depicted in Fig. 2.

The biologically-motivated rearrangement operations we consider include inversions (implying as well change of orientation) of chromosomal segments containing one or more markers, reciprocal translocations (of telomere-containing segments – suffixes or prefixes – of two chromosomes) and chromosome fission or fusion. We use the versatile rearrangement algorithm of Bergeron *et al.* [1], which we constrain to allow only the operations we have listed.

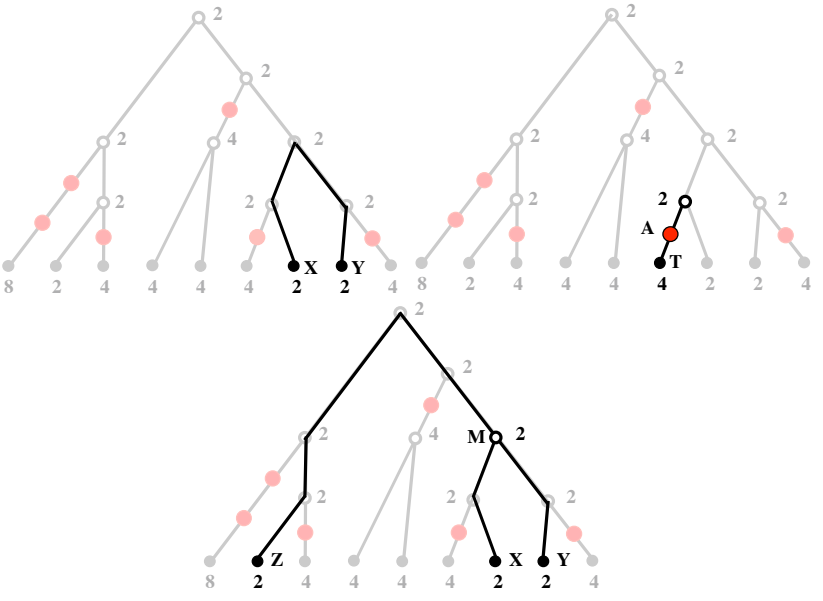

Fig. 2. Clockwise, from upper left: Genomic distance, Genome halving, Rearrangement median

Genome halving. Given a genome T containing a set of markers, each of which appears twice on the genome, on the same or on different chromosomes, how can we construct a genome A containing only one copy of each marker, and such that the genome $A \oplus A$ consisting of two copies of each chromosome in A minimizes $d(T, A \oplus A)$? This problem is illustrated in the rightmost example in Fig. 2. Here we use a linear-time algorithm for solving this problem [6].

Rearrangement median. Given three genomes X,Y and Z, how can we find the *median* genome M such that $d(X, M)+d(Y, M)+d(Z, M)$ is minimized. For this NP-hard problem, illustrated in the bottom example in Fig. 2. we implement a heuristic using the principles of Bourque's MGR [2], but based on the constrained version of the Bergeron *et al.* [1] algorithm.

4 Parts Already in Place

In this section we discuss heuristics for prototypical phylogeny problems involving doubling descendant, and either one or two related unduplicated genomes.

Let T be a doubling descendant, i.e., with n different chromosomes, and $2m$ markers, $g_{1,1} \cdots, g_{1,m}; g_{2,1}, \cdots, g_{2,m}$, dispersed in any order on these chromosomes. For each i, we call $g_{1,i}$ and $g_{2,i}$ "duplicates", and the subscript "1" or "2" is assigned arbitrarily. A potential ancestral doubled genome of T is written $A \oplus A$, and consists of $2n'$ chromosomes, where some half (n') of the chromosomes

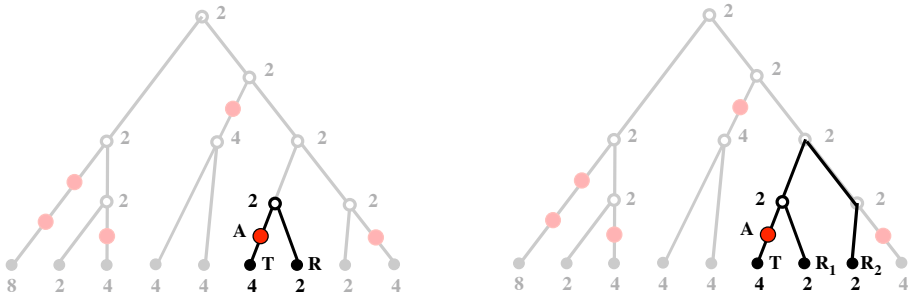

Fig. 3. Genome halving with one (left) or two (right) unduplicated outgroups

contains exactly one of each of $g_{1,i}$ or $g_{2,i}$ for each $i = 1, \cdots, m$. The remaining n' chromosomes are each identical to one in the first half, in that where $g_{1,i}$ appears on a chromosome in the first half, $g_{2,i}$ appears on the corresponding chromosome in the second half, and *vice versa*. We define A to be either of the two halves of $A \oplus A$, where the subscript 1 or 2 is suppressed from each $g_{1,i}$ or $g_{2,i}$. These n' chromosomes, and the m markers they contain, $g_1, \cdots, g_m$, constitute a potential ancestor of T that incurred the doubling event .

Genome halving with an outgroup. With reference to the left of Fig. 3, consider T and and a related unduplicated genome R with markers orthologous to $g_1, \cdots, g_m$. Our problem is to find an unduplicated genome A that minimizes

$$D(T, R) = d(R, A) + d(A \oplus A, T). \tag{1}$$

Our solution in [11], as on the left of Fig. 4, is to generate the set $\mathbf{S}$ of genome halving solutions, then to focus of the subset $X \in \mathbf{S}' \subset \mathbf{S}$ where $d(R, X)$ is minimized. We then minimize $D(T, R)$ by seeking heuristically for A along any trajectory between elements of $\mathbf{S}'$ and the outgroups.

Genome halving with two outgroups. With reference to the right of Fig. 3, consider T and two unduplicated genomes R_1 and R_2 with markers orthologous to $g_1, \cdots, g_m$. Our problem here is to find a diploid genome A and a median genome M of A, R_1 and R_2 that minimize

$$D(T, R_1, R_2) = d(R_1, M) + d(R_2, M) + d(A, M) + d(A \oplus A, T). \tag{2}$$

Our solution in [9], as on the right of Fig. 4, is to generate the set $\mathbf{S}$ of solutions of the genome halving problem, then to focus of the subset $X \in \mathbf{S}' \subset \mathbf{S}$ where $d(R_1, M) + d(R_2, M) + d(X, M)$ is minimized. Then the A minimizing $D(T, R_1, R_2)$ is sought, heuristically, along all trajectories between all elements $X \in \mathbf{S}'$ and $M(X)$.

5 The Case of Two Doubling Descendants

Two related doubled descendants may arise in two ways, depending on the timing of the speciation event in relation to the doubling. Either speciation at V follows a single doubling event, as at A on the left of Fig. 5, or the speciation precedes two independent doubling events in the two lineages, as at A and B on the right of the figure. Knowing which of the two scenarios is correct depends on knowing whether their common ancestor is doubled or not, information obtained from the algorithm in Section 2 or other data.

We will introduce new methods based on tweaking the distance and halving algorithms, conserving the optimality of the solutions, but allowing one of them to affect the arbitrary choices required to construct the solution for the other. First we sketch the halving algorithm.

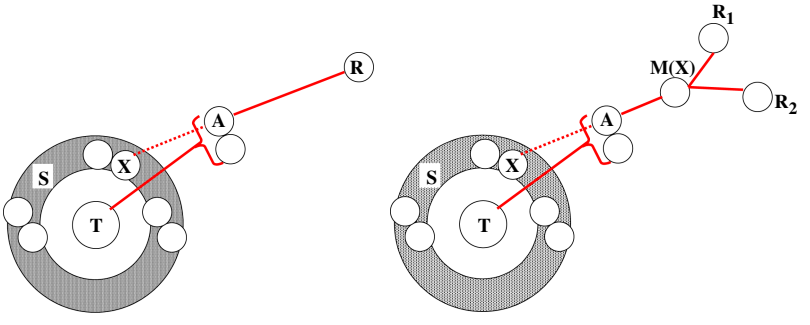

Fig. 4. Halving a doubling descendent T, with one (R) or two (R_1, R_2) unduplicated outgroups. The double circles represent two copies of potential ancestral genomes, including solutions to the genome halving in **S**, and those on best trajectories between **S** and outgroups.

5.1 Halving

Without entering into all its details, we can present enough of the essentials of the halving algorithm to understand the techniques we use in our heuristics.

As a first step each marker x in a doubled descendant is replaced by a pair of vertices (x_t, x_h) or (x_h, x_t) depending if the DNA is read from left to right or right to left. The duplicate of marker $x = (x_t, x_h)$ is written $\bar{x} = (\bar{x_t}, \bar{x_h})$. Of course $\bar{\bar{a}} = a$.

Following this, for each pair of neighbouring markers, say (x_t, x_h) and (y_h, y_t), the two adjacent vertices x_h and y_h are linked by a black edge, denoted $\{x_h, y_h\}$ in the notation of [1]. For a vertex at the end of a chromosome, say y_t, it generates a virtual edge of form $\{y_t, O\}$.

The edges thus constructed are then partitioned into *natural graphs* according to the following principle: If an edge $\{a, b\}$ belongs to a natural graph, then so does some edge of form $\{\bar{a}, c\}$ and some edge of form $\{\bar{b}, d\}$. If a natural graph has an even number of edges, it can be shown that in all optimal ancestral

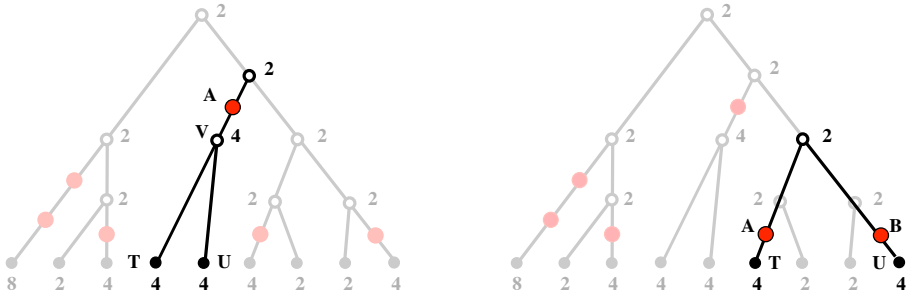

Fig. 5. Left: Doubling, then speciation. Right: Speciation, then two independent doublings.

doubled genomes, the edges coloured gray, say, representing adjacent vertices in the ancestor, and incident to one of the vertices in this natural graph, necessarily have as their other endpoint another vertex within the same natural graph[1].

For all other natural graphs, there are one or more ways of grouping them pairwise into *supernatural graphs* so that an optimal doubled ancestor exists such that the edges coloured gray incident to any of the vertices in a supernatural graph have as their other endpoint another vertex within the same supernatural graph.

Along with the multiplicity of solutions caused by different possible constructions of supernatural graphs, within such graphs and within the natural graphs, there may be many ways of drawing the gray edges. Without repeating here the lengthy details of the halving algorithm, it suffices to note that these alternate ways can be generated by choosing one of the vertices within each supernatural graph as a starting point.

5.2 Doubling First

Given two doubling descendants T and U as on the left of Figure 5, we would ideally like to find the doubling descendant V that minimizes $d(T,V)+d(V,U)+d(V,A\oplus A)$, where A is any solution of the halving problem on V. Though d is calculated in linear time, multiple genome rearrangement problems based on d (e.g., the median problem in Section 3) are hard, so here we propose a somewhat constrained version of our problem, where V is assumed to be on a shortest trajectory between T and U. Because $d(T,V)+d(V,U)=d(T,U)$ is then constant, the problem becomes that of finding V to minimize $d(V,A\oplus A)$.

Because it is an edit distance, a genomic distance measurement $d(T,U)$ is associated with at least one trajectory containing $d(T,U)-1$ genomes as well as

[1] Space precludes us from elaborating on the connection between the optimality criterion – the minimum number of rearrangements to transform the doubled ancestor to the doubled descendant – and the nature of the bicoloured graph defined by the black and gray edges. Suffice it to indicate that this involves maximizing the number of (alternating coloured) cycles and certain paths that make up this graph.

T and U themselves, where each successive pair of genomes along the trajectory differ by exactly one rearrangement operation.

Before explaining a heuristic search for a solution to the constrained version of the problem, we recall the edge notation we use to represent the adjacencies in a genome [1]. If two vertices a and b from different markers are adjacent in a genome, we represent this by an edge $\{a, b\} = \{b, a\}$; for a vertex c is at the end of a chromosome and hence adjacent to no other vertex, we construct a virtual edge $\{c, O\}$. Then any rearrangement operation can be represented by an operation on one or two terms in the representation, such as $\{a, b\}, \{c, d\} \rightarrow \{b, d\}, \{a, c\}$ or $\{a, b\} \rightarrow \{b, O\}, \{a, O\}$ or $\{a, b\}, \{c, O\} \rightarrow \{b, O\}, \{a, c\}$.

We initialize $T^* = T, U^* = U$. Then our heuristic consists of a search, at each step, for the "most promising" operation that moves T^* towards U^* or U^* towards T^*. For each operation, we define a score $W = x + 6y$ as follows. The y component, which is heavily weighted, measures whether the operation actually diminishes $d(V, A \oplus A)$, while the x measures whether the operation only increases the potential of diminishing $d(V, A \oplus A)$ in a subsequent operation.

Consider the possible operations that remain on a trajectory from T to U, i.e., if V_1 is transformed into V_2 by the operation, then $d(T, V_2) = d(T, V_1) + 1$ and $d(V_2, U) = d(V_1, U) - 1$. We set $y = d(V_1, A_1 \oplus A_1) - d(V_2, A_2 \oplus A_2) + 1$, where A_1 and A_2 are solutions of the halving problem for V_1 and V_2, respectively.

In evaluating an operation changing T^*, such as $\{a, b\}, \{c, d\} \rightarrow \{b, d\}, \{a, c\}$, we consider the following eight pairs: $\{a, b\}$,$\{c, d\}$, $\{b, d\}$,$\{a, c\}$,$\{\bar{a}, \bar{b}\}$,$\{\bar{c}$,$\bar{d}\}$, $\{\bar{b}, \bar{d}\}$, $\{\bar{a}, \bar{c}\}$.

The operation would clearly seem advantageous for subsequent operations if $\{\bar{b}, \bar{d}\}$ and/or $\{\bar{a}, \bar{c}\}$ were in T^* and/or U^*. There are from zero to four advantageous possibilities. In addition, although one of $\{b, d\}, \{a, c\}$ must be in U^* for the operation not to veer from an optimal trajectory, it is not necessary that both of them be. There are zero or one advantageous possibilities. We count how many h of the total of five advantageous possibilities occur and set $x = h + 1$.

The score W is in the range $[1, 18]$. We calculate W_{T^*} in this way and W_{U^*} by considering operations changing U^* in the direction of T^*. Let $W_X = \max_{\text{all operations}} W_{X^*}$.

If $W_T \geq W_U$ and $W_T \geq 6$, we apply the highest score operation to T^*. Otherwise apply the highest score operation to U^*, as long as this $W_U > 1$. The results of this operation and any other having the same score are added as nodes to a search tree. (The search tree was initialized when $T^* = T$ and $U^* = U$.) When there are no more operations that can be applied, we continue to build the search tree at a higher node. Finally, the leaves of the search tree are examined to find the highest scoring genome to be V, the last common ancestor of T and U.

Using a range $W \in [1, 18]$ proves clearly better than simply choosing evaluating an operation according to whether it $y = 1$ or $y \neq 1$. For example, in simulations generated with $d(T, V) = 60, d(V, U) = 55, d(V, A \oplus A) = 24$, the average estimate $d(V, A \oplus A)$ using an 18-value scale was 29.8, an overestimate of 24%, compared to 31.7 with a two-value scale, an overestimate of 32%.

5.3 Speciation First

In Section 5.2, $d(T, V) + d(V, U)$ was fixed and the problem was to find the common ancestor V with the shortest history from the doubling event. We now consider the halving distances of T and U both to be fixed, and look for the particular unduplicated genomes, ancestral to T and U, that are closest together. Our Algorithm 1 simultaneously halves T and U, choosing the initial vertex within each of the supernatural graphs (henceforward SNGs) so as to maximize the number of gray edges in common in the two ancestral genomes being constructed.

Both this heuristic and the one in Section 5.2 are basically $O(m^3)$ to arrive at a single estimate. This, however, generally produces a locally optimal solution. This is improved by maintaining a search tree in association with each algorithm. Then the running time is controlled by how large a search tree is maintained in the quest for lower estimates.

6 Simulations

Simulations of the doubling first model (five chromosomes, number of markers $m = 200$, inversions to translocations proportion 5:3, random choice of chromosomes to be rearranged, random breakpoints on chromosomes) show that our algorithm accurately reconstructs the number ν of rearrangements (ten replications for each value of ν) between the doubling event and the speciation event, as long as this is not too large (Fig. 6, top). With a longer interval between doubling and speciation, the halving algorithm reconstructs the unduplicated ancestor too economically. This, however, is a function not of the number of rearrangements in the simulation, but of the number of markers. If the number of markers is doubled from 200 to 400, the inferred number of rearrangements is corrected, as indicated by the square dot in the figure.

Simulations of the speciation first model ($m = 400$) show that while the genome halving distances accurately estimate the number of rearrangements between doubled ancestor and doubled descendant in the simulation (data not shown here), the estimated unduplicated ancestors are further apart than the genomes actually generated in the simulation (Fig. 6, bottom). This bias increases dramatically as a function, not of the distance itself, but of the amount of rearrangement these ancestors incur to produce the observed doubling descendant. When this "age" is 20, 50 and 80 rearrangements, the bias in the distance between the ancestors increases from 4 to 18 to 37, respectively. This reflects the severely non-unique result of the halving algorithm, which our algorithm attenuates by forcing the reconstructed doubled genomes to resemble each other as much as possible, but cannot eliminate, especially as the age of the doubling events recedes into the past.

Nonetheless, the superiority of our algorithm in constraining the two simultaneous halving processes to create ancestor genomes as close as possible, in comparison with a search over all pairs in $\mathbf{S_T} \times \mathbf{S_U}$, the Cartesian product of the two complete sets of solutions of the halving algorithm, is clear in another experiment.

Algorithm 1

Construct σ_T and σ_U, the set of supernatural graphs for T and U, respectively.
Initialize $\sigma_T^{(1)}$ = the subset of SNGs with 2 black edges and $\sigma_T^{(0)} = \sigma_T \setminus \sigma_T^{(1)}$
Initialize $\sigma_U^{(1)}$ = the subset of SNGs with 2 black edges and $\sigma_U^{(0)} = \sigma_U \setminus \sigma_U^{(1)}$
Step1: Order σ_T **and** σ_U
while there remain SNGs in $\sigma_T^{(0)}$ or SNGs in $\sigma_U^{(0)}$
  while there remain SNGs in $\sigma_T^{(0)}$ and either $\sigma_U^{(0)}$ is empty or the number of black edges in $\sigma_T^{(1)}$ is no more than in $\sigma_U^{(1)}$, we find a SNG in $\sigma_T^{(0)}$, to move from $\sigma_T^{(0)}$ to $\sigma_T^{(1)}$, as follows:
    for each SNG s in $\sigma_T^{(0)}$, to count the maximum possible number of gray edges it could have in common with SNGs in $\sigma_U^{(1)}$:
      for $i = 1, \cdots, |\sigma_U^{(1)}|$, if SNG s has k_i vertices in common with t_i, the $i-$th SNG in $\sigma_U^{(1)}$, the maximum number of gray edges they have in common is $[\frac{k_i}{2}]$.
      Then the score of s is $\sum_{i,\cdots,|\sigma_U^{(1)}|}[\frac{k_i}{2}]$.
    We add the highest scoring s to $\sigma_T^{(1)}$.
  end while
  while there remain SNGs in $\sigma_U^{(0)}$, and either $\sigma_T^{(0)}$ is empty or the number of black edges in $\sigma_U^{(1)}$ is less than $\sigma_T^{(1)}$), we find a SNG in $\sigma_U^{(0)}$, to move from $\sigma_U^{(0)}$ to $\sigma_U^{(1)}$, in the analogous way as for T
  end while
end while
Step2: Adding gray edges to σ_T **and** σ_U
For the root node of the search tree, add gray edges to all 2-edge SNGs in σ_T and σ_U
while there remain SNGs in σ_T or σ_U without gray edges.
  while there remain SNGs in σ_T without gray edges and either all SNGs in σ_U have gray edges or the number of gray edges in σ_T is no more than the number of gray edges in σ_U, let s be the first SNG in σ_T (according to the order in which it was added to $\sigma_T^{(1)}$) that has no gray edges. If s has l black edges, then we have l ways to choose the first black edge in this s, and 2 choices for orienting this edge, $2l$ choices in all, after which the dedouble algorithm proceeds deterministically to add gray edges within the SNG s.
  We add nodes to the search tree representing all the choices (out of the $2l$) that maximize the number of gray edges in common with σ_U.
  end while
  while there remain SNGs in σ_U without gray edges either all SNGs in σ_T have gray edges or the number of gray edges in σ_U is less than the number of gray edges in σ_T, let s be the first SNG in σ_U(according to the order in which it was added to $\sigma_U^{(1)}$) that has no gray edges. We use the same process as with T to get the best orderings within s and the associated gray edges.
  end while
end while
Solutions to the genome halving can then be found by tracing backwards from any leaf in the search tree.

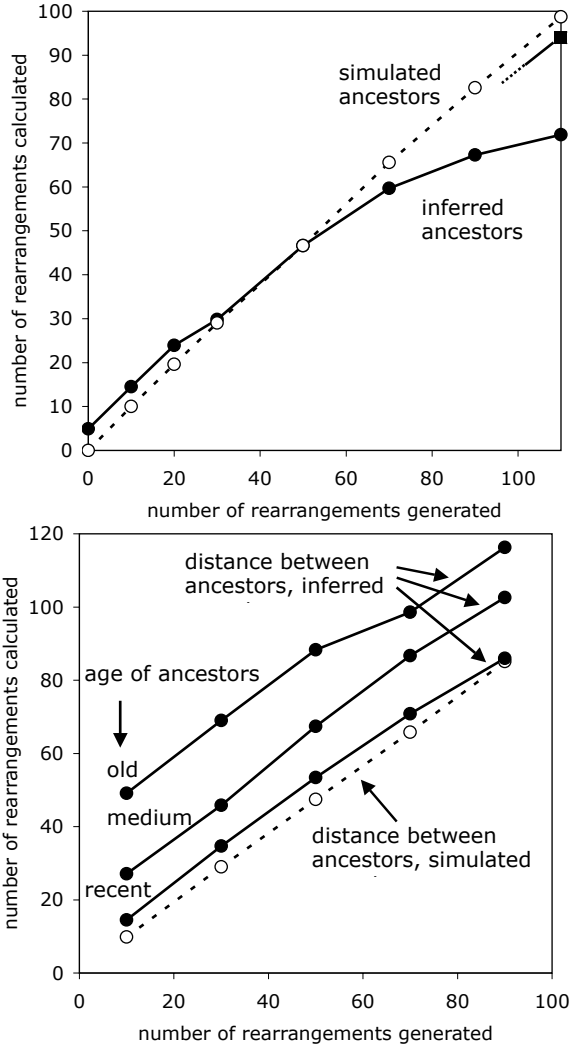

Fig. 6. Estimated distance: top, between doubling and speciation (age of ancestor=50), bottom, between unduplicated ancestors (ages: old=80, medium=50, young=20)

We set the initial number of markers to be 150, randomly assigned to 8 chromosomes. Then we carried out 45 random rearrangements to create one doubling ancestor and 38 independent rearrangements to create another. After tetraploidization formed two 300-marker genomes, we applied another 42 and 50 rearrangements, respectively, to create the modern doubling descendants. Then, using our knowledge of the ancestral genomes, we found that the distance between the two simulated ancestors was 75 and that the halving distances were 38 and 50, respectively. Using our speciation first algorithms on the two doubling

descendants, we reached an inter-ancestor distance $d(A, B) = 84$ (instead of the simulated distance of 75) after three hours of calculation while the search of the Cartesian product only dropped to 87 (from 102) after 24 hours of calculation, involving almost 1,000,000 pairs of optimal ancestors.

7 Genome Doubling in Yeast

Wolfe and Shields [10] discovered an ancient genome doubling in the ancestry of *Saccharomyces cerevisiae* in 1997 after this organism became the first to have its genome sequenced [7]. According to [8], the recently sequenced *Candida glabrata* [4] shares this doubled ancestor. We extracted data from YGOB (Yeast Genome Browser) [3], on the orders and orientation of the exactly 600 genes identified as duplicates in both genomes, i.e., 300 duplicated genes.

We were able to obtain information from YGOB about which of the two duplicates in one genome is orthologous to which duplicate in the other genome. This is essential to the algorithm in Section 5.2. In general, we would have to infer this information through sequence comparison methods. This question is not pertinent to the algorithm in Section 5.3.

Though the results of the algorithm in Section 2 suggests that the theory in [8] is the most parsimonious, there is still enough uncertainty in yeast phylogenetics and enough independent occurrences of genome doubling, that it is worth comparing the results of our two methods to dispute or confirm the common doubled ancestor hypothesis. In Fig. 5 we compare the analysis in the left hand diagram with that in the right, on the yeast data and on data of approximately the same size generated first according to the doubling first model and then according to speciation first.

We first analyzed the yeast data using the doubling first and speciation first algorithms. The results appear in the centre row of Table 1. (Because of the asymmetry of the doubling first algorithm with respect to T and U, there are two sets of inferences for this case.) We then used the numbers of rearrangements inferred for yeast, using the same number of markers and chromosomes, to simulate the same number of rearrangements in a random model, both with doubling first and speciation first.

We then applied both algorithms, doubling first and speciation first, to both sets of data. Note first in Table 1 that the number of rearrangements inferred for the doubling first model using the doubling first algorithm is not exactly the same as that used to generate the data, and likewise for the speciation first case. This is normal, because the inference of rearrangements often is more economical than the rearrangements actually used.

The rows in Table 1 show that the doubling first analysis is better than the speciation first analysis (457 rearrangements versus 632) when the data are generated by doubling first, whereas the speciation first analysis is better (589 versus 604) when the data is generated with speciation first. The doubling first analysis clearly accounts better for the yeast data (505-521 versus 622), while the simulations assure that the biases in the two methods cannot be invoked, so our analysis confirms the hypothesis in [8].

Table 1. Doubling first (d.f) and speciation first (s.f.) analyses each produce a more parsimonious analysis of simulations produced by the corresponding model (d.f. or s.f., respectively). Averages of at least five simulations shown, but the effect holds for each simulation individually. The d.f. analysis gives a far better fit to the yeast data than s.f. Second yeast row reverses the roles of U and T in the algorithm.

analysis→	doubling first (d.f.)				speciation first (s.f.)			
data source↓	$d(T,V)$	$d(V,U)$	$d(V,A\oplus A)$	total	$d(T,A\oplus A)$	$d(A,B)$	$d(U,B\oplus B)$	total
sim by d.f.:	102	213	166	481				
inferred:	119	181	157	457	214	163	255	632
yeast:	92	245	168	505	193	179	250	622
	122	215	184	521				
sim by s.f.:					177	164	225	566
inferred:	146	354	104	604	164	228	197	589

8 Conclusions

Our previous work on integrating genome halving and other algorithms as a way of incorporating polyploids into rearrangement phylogeny used this software "off the shelf", searching all the many alternate outputs from one as inputs to the other. In the present paper we have avoided an exhaustive search strategy by intervening at the choice points in the genomic distance algorithm in the case of the doubling first problem and in the genome halving algorithm in the case of the speciation first problems. We have shown that these heuristics increase the efficiency of the search and to provide better upper bounds.

The main difficulty in this problem area remains the great multiplicity of solutions to the halving problem. Though this was only encountered here in the speciation first problem, leading to a overestimation of the inter-ancestor distance, it will also have to be dealt with in the doubling first scenario, when the inferred ancestor has to be integrated into a larger phylogenetic tree and compared to other doubled or unduplicated genomes, as in [11] and [9].

References

1. Bergeron, A., Mixtacki, J., Stoye, J.: A unifying view of genome rearrangements. In: Bücher, P., Moret, B.M.E. (eds.) WABI 2006. LNCS (LNBI), vol. 4175, pp. 163–173. Springer, Heidelberg (2006)
2. Bourque, G., Pevzner, P.: Genome-scale evolution: Reconstructing gene orders in the ancestral species. Genome Research 12, 26–36 (2002)
3. Byrne, K.P., Wolfe, K.H.: The Yeast Gene Order Browser: combining curated homology and syntenic context reveals gene fate in polyploid species. Genome Research 15, 1456–1461 (2005)
4. Dujon, B., Sherman, D., Fischer, G., et al.: Genome evolution in yeasts. Nature 430, 35–44 (2004)
5. El-Mabrouk, N., Sankoff, D.: Hybridization and genome rearrangement. In: Crochemore, M., Paterson, M.S. (eds.) Combinatorial Pattern Matching. LNCS, vol. 1645, pp. 78–87. Springer, Heidelberg (1999)

6. El-Mabrouk, N., Sankoff, D.: The reconstruction of doubled genomes. SIAM Journal on Computing 32, 754–792 (2003)
7. Goffeau, A., Barrell, B.G., Bussey, H., et al.: Life with 6000 genes. Science 275, 1051–1052 (1996)
8. Kurtzman, C.P., Robnett, C.J.: Phylogenetic relationships among yeasts of the Saccharomyces complex determined from multigene sequence analyses. FEMS Yeast Research 3, 417–432 (2003)
9. Sankoff, D., Zheng, C., Zhu, Q.: Polyploids, genome halving and phylogeny. Accepted for ISMB (2007)
10. Wolfe, K.H., Shields, D.C.: Molecular evidence for an ancient duplication of the entire yeast genome. Nature 387, 708–713 (1997)
11. Zheng, C., Zhu, Q., Sankoff, D.: Genome halving with an outgroup. Evolutionary Bioinformatics 2, 319–326 (2006)