BIOINFORMATICS

Vol. 23 ISMB/ECCB 2007, pages i433–i439
doi:10.1093/bioinformatics/btm169

Polyploids, genome halving and phylogeny

David Sankoff[1,*], Chunfang Zheng[2] and Qian Zhu[3]

[1]Department of Mathematics and Statistics, [2]Department of Biology and [3]Department of Biochemistry, University of Ottawa, Ottawa K1N 6N5, Canada

ABSTRACT

Motivation: Autopolyploidization and allopolyploidization events multiply the number of chromosomes and genomic content. Genome rearrangement phylogenetics requires that all genomes analyzed have the same set of orthologs, so that it is not possible to include diploid and polyploid genomes in the same phylogeny.

Results: We propose a framework for solving this difficulty by integrating the rearrangement median and genome halving algorithms. Though the framework is general, some problems remain open. We implement a heuristic solution to the prototypical case of a tree with one tetraploid and two diploid genomes, and apply it to study the evolution of cereals and of yeast.

Contact: sankoff@uottawa.ca

1 INTRODUCTION

Phylogenomics based on cross-species comparisons of synteny block order (henceforward *rearrangement* phylogenetics) provides an approach to phylogenetics independent of that based on nucleotide or amino acid sequence divergence. The order-based approach takes advantage of the periodic and cumulative rearrangement of genomic material by evolutionary processes, such as inversion, reciprocal translocation and transposition. The basic methods require that the genomic content be roughly the same in all the organisms being compared, so that every 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 deletion, insertion and duplication of segments.

Many genomes have been shown to result from an ancestral doubling, or tetraploidization, event, after which meiosis is characterized not by the normal pairings of one maternal and one paternal chromosome, but by quadrivalent alignment of chromosomes or other combinations. Tetraploidization is followed by a period of re-diploidization, where distinct pairings again emerge, though in twice the original number, a process mediated by sequence divergence and by genome rearrangement through intra- and interchromosomal movement of genetic material. The present-day genome (often still referred to loosely as a tetraploid) can be decomposed into a set of duplicated synteny blocks dispersed among the chromosomes. There is usually no obvious way of partitioning the blocks into two sets according to which ones were together in the original tetraploid.

Rearrangement phylogeny algorithms are not applicable since there is a two-to-one relationship between blocks in the former tetraploid and those in related diploid species, whereas these algorithms require a one-to-one correspondence.

Tetraploidization may also occur as a fusion of two distinct but related genomes (allotetraploidy) instead of the doubling of a genome (autotetraploidy), and both types of polyploidization may recur during evolution, so that instead of a $2n$ diploid number, the descendant (polyploid) genome will have $2rn$, where $r > 1$.[1] These genomes will be constituted not by duplicated blocks, but by a set of blocks with r homologous copies each, dispersed among the chromosomes.

In this article, we provide an overall strategy for rearrangement phylogeny for sets of related genomes that include some that have undergone polyploidization, including allopolyploidization. We specifically attack the 'small' phylogenetic problem, i.e. identifying the ancestral genomes for a given phylogeny that jointly minimize the sum of the rearrangement distances along the branches of that phylogeny. To take into account allopolyploidy, the phylogeny must be reticulated.

In Section 2, we outline a model for generating an arbitrary pattern of polyploidy observed at the tips of a reticulate phylogeny. Based on this model, we then present an algorithm for inferring the ploidy of the ancestral genomes in terms of an economical set of autopolyploidization and allopolyploidization events along the edges of the phylogeny graph. Once we have the ancestral ploidies, we can approach the actual rearrangement problem. We identify three kinds of component of this problem, 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 'de-ploidization' calculation for inferring the genome of an ancestral polyploid based on internal evidence from its modern descendant only and the third a 'medianizing' process for inferring an ancestral genome from its three neighboring genomes in a binary branching tree. In Section 3, we show how to integrate algorithms for the three components into an overall procedure for inferring the ancestral polyploids in a given phylogeny, and we describe in particular detail the prototypical case of one tetraploid and two related diploids. In Sections 4 and 5, we apply our method to a small data set on maize and a large data set on yeast, respectively.

*To whom correspondence should be addressed.

[1] Genomes with odd ploidy are generally deemed to be infertile because of the impossibility of segregating into haploids containing equal numbers of chromosomes during meiosis.

2 MODEL, INFERENCE AND DECOMPOSITIONS

The simplified assumptions we will adopt in this abstract are that polyploidization occurs either by tetraploidization of a genome, namely replacing each of its chromosomes by two identical chromosomes, so the diploid number goes from $2n$ to $4n$, or by the fusion of two different genomes of diploid numbers $2n$ and $2m$, respectively, merging the two sets of chromosomes, and producing a $2(n+m)$ allopolyploid. Following the polyploidization, the genome evolves via inversion of chromosomal segments, reciprocal translocation between two chromosomes, or chromosome fusion and fission, and may further polyploidize at any time.

We will assume the evolutionary histories to be binary branching trees, with allopolyploidy events represented by horizontal reticulations between branches of the tree, as illustrated in Figure 1. The model imposes the equations in the illustration: each autopolyploid must have ploidy equal to a non-negative exponent of 2, times the ploidy of its immediate ancestor. Each allopolyploid must have ploidy equal to the sum of its contributing genomes. The allopolyploidy events are given, though not the ploidy of the participating genomes, which must be inferred, and the autopolyploidy events are to be inferred.

This model is simplified and cannot account for all possible observations of even-numbered ploidies at the leaves of the phylogeny; a full model of polyploidy in phylogenetic context would allow for events such as the fusions of a polyploid with an earlier diploid version of itself. Such a model, worked out in the full version of this article, can account for all possible observations of even-numbered ploidies at the leaves of the phylogeny, but can also give rise to a great multiplicity of solutions.

Because our restricted version of this problem here does not generate all possible combinations of observations at the leaves of the tree, the solution to the ancestral ploidy assignment problem does not always exist for an arbitrary data set of present-day ploidies. When it does exist, it can be obtained by solving a system of equations such as that in Figure 1, with the objective of minimizing the sum of the exponents in the autopolyploidization equations. Generally, the ploidy of the root is as high as possible, consistent with a minimum of autopolyploidization events along all the branches.

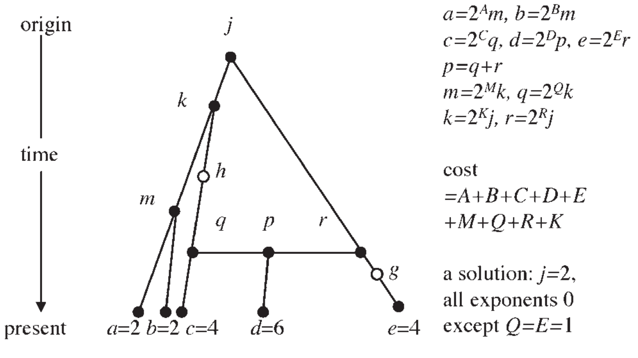

Fig. 1. Example of ploidy inference problem. Genomes labeled by ploidies, observed only for leaves of phylogeny. Tetraploidy events inferred at g and h, or alternatively on the branches jr and qc.

Once we have inferred the ploidy of the ancestral genomes, how are we to approach our original problem: to reconstruct the synteny block order of the ancestral genomes and thus infer the cost of the phylogeny in terms of rearrangement events? Elements of the solution are discussed in Section 3.1.1 below. The first point to stress is that the **rearrangement distance** can only be directly calculated between two genomes that have a common polyploidization history. Thus, we can calculate the rearrangement distance between the genomes labeled a and b in Figure 1, but not between a and c. What is required is to take account of the inferred transition from diploid to tetraploid, the autopolyploidization event h, on the path between q and k. We add the distance between the tetraploids at h and c to the distance between the diploids at h and a. To be able to do this, we first find the synteny block order at h using the **genome halving** algorithm.

We may further ask, even if we can calculate h, how can we know the synteny block order for an ancestor like that labeled m in Figure 1? This requires a **median** algorithm. Other questions to be answered before all kinds of ancestral genomes can be inferred, and the total branch length of the phylogeny evaluated, are listed in Section 3.1.4.

3 THE ALGORITHMS

In this section, we discuss a local search heuristic for the solution to a prototypical phylogeny problem involving one genome descended from a tetraploid and two related diploids. The main focus of this work is to produce an accurate initialization. It is based on integrating three existing algorithms, which we can only cite in this abstract.

3.1 Existing and missing resources

3.1.1 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. Each genome is composed of a (possibly different) number of chromosomes containing linearly ordered terms. Comparison of the two genomes induces a decomposition of each into a set of synteny blocks. The set of blocks is the same for each genome, but it is differently partitioned among the chromosomes, differently ordered within the chromosomes, and the left-right orientation of a block may also differ in the two genomes.

The biologically motivated rearrangement operations we consider include inversions (implying as well change of orientation) of chromosomal segments containing one or more blocks, reciprocal translocations (of telomere-containing segments—suffixes or prefixes—of two chromosomes) and chromosome fission or fusion. Here we make use of a versatile rearrangement algorithm recently introduced by Bergeron *et al.* (2006), which we constrain to allow only the operations we have listed.

3.1.2 Genome halving Given a genome T that can be decomposed into a set of synteny blocks, 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 block, and such that the genome $A \oplus A$ consisting of two copies of each chromosome in A minimizes $d(T, A \oplus A)$? Here we use the linear-time algorithm for solving this problem due to El-Mabrouk and Sankoff (2003).

3.1.3 Rearrangements 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, we implement a heuristic using the principles of the [Bourque and Pevzner, 2002] MGR (multiple genome rearrangement) algorithm, but based on the constrained version of the Bergeron *et al.* (2006) distance algorithm.

3.1.4 Open questions To fully solve the inference problem as stated, even within the limitations imposed by the heuristic implementation of the median problem and the heuristic steps in the main algorithm in Section 3.2 below, we would have to generalize the genome halving problem in several directions:

- Given two tetraploid (or 2^α-ploid) genomes X and Y, i.e. with two (or $2^{\alpha-1}$) copies each of every syntenic block, find the matching of each pair (or set of $2^{\alpha-1}$) paralogs between the two genomes that minimizes the rearrangement distance.
- Given a genome P with ploidy $2p = 2(r + s), r, s > 0$, find the $2r$-ploid and $2s$-ploid genomes R and S, respectively, such that the distance $d(P, R \oplus S)$ is minimized.
- Given a genome Q with ploidy $2^\alpha, \alpha > 1$, find the $2^{\alpha-1}$-ploid A such that the distance $d(Q, A \oplus A)$ is minimized.

3.2 Strategy for the problem of one tetraploid and two diploids

Let T be a genome with diploid number $4n$, i.e. $2n$ pairs of (identically ordered) maternal and paternal chromosomes, and $2m$ syntenic blocks, $g_{1,1} \ldots, g_{1,m}; g_{2,1}, \ldots, g_{2,m}$, dispersed in any order on the $2n$ different 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 tetraploid' of T is written $A \oplus A$, and consists of $2n'$ chromosomes, where some half (n') of the chromosomes contains exactly one of each of $g_{1,i}$ or $g_{2,i}$ for each $i = 1, \ldots, m$. 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 syntenic blocks they contain, $g_1, \ldots, g_m$, constitute a potential 'ancestral diploid' of T.

A solution of the genome halving problem for T is any A such that $d(A \oplus A, T)$ is minimal. There are generally many different solutions to this problem.

Consider an unrooted tree with three leaves, T and two diploid genomes R_1 and R_2 with blocks orthologous to $g_1, \ldots, g_m$, as in Figure 2a. Our central problem is to find a diploid genome A and a median genome M of A, R_1 and R_2 that minimize

$$D(T, R_1, R_2) = d(R_1, M) + d(R_2, M) + d(A, M) + d(A \oplus A, T). \qquad (1)$$

There is no requirement that A be a solution to the genome halving problem, but since they already minimize one term of D, some of these solutions might be good initial guesses for an optimal A. Let $\mathbf{S}$ be the set of solutions of the genome halving algorithm for T. Initially in our heuristic, schematized in Figure 2b, we confine our search to $\mathbf{S}$.

For each solution $X \in \mathbf{S}$, we calculate the median distance $d(R_1, M(X)) + d(R_2, M(X)) + d(X, M(X))$, as in Figure 2c. This is the bottleneck in our procedure, since $\mathbf{S}$ may be very large, and an accurate calculation of the median is costly for each element of $\mathbf{S}$. When the number of markers m is small, say a few dozen, as to be illustrated in Section 4 below, it is possible to do evaluation of $\mathbf{S}$. When m is in the hundreds, as to be illustrated in Section 5 below, we resort to a random sample of the genomes in $\mathbf{S}$.

We then define

$$\mathbf{S}' = \{X \in \mathbf{S} | d(R_1, M(X)) + d(R_2, M(X)) + d(X, M(X)) \text{ is a minimum}\}. \qquad (2)$$

By definition, there is no minimizing genome in $\mathbf{S} \backslash \mathbf{S}'$. To look for a minimizing A outside of $\mathbf{S}$, we first guess that any such genome will be found on a path between some element $X \in \mathbf{S}'$ and $M(X)$, as in Figure 2d. We calculate the $d(X, M(X))$ genomes, other than X, on a parsimonious trajectory $X, X^{(1)}, X^{(2)}, \cdots, M(X)$ from X to $M(X)$. Note that $d(X^{(i)}, M(X)) = d(X, M(X)) - i$. Then we search for an $X^{(i)}$ such that

$$d(X^{(i)}, M(X)) + d(X^{(i)} \oplus X^{(i)}, T) < d(X, M(X)) + d(X \oplus X, T). \qquad (3)$$

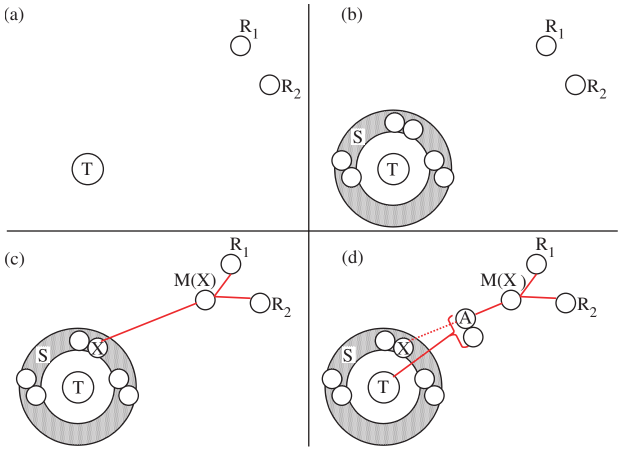

Fig. 2. Strategy for phylogenetically constrained genome halving. (**a**) Descendant T of ancestral tetraploid, with two related diploids R_1 and R_2. (**b**) Set $\mathbf{S}$ of solutions of genome halving of T, showing pairs of fused identical diploids. (**c**) Solution $X \in \mathbf{S}$ that also induces minimizing solution $M(X)$ of the median problem on X, R_1 and R_2. (**d**) Genome A minimizing objective function among all genomes on any trajectory between X and $M(X)$.

For relatively small examples, e.g. for the data in Section 4, we can also iterate on the median step, and look for

$$\begin{aligned} d(X^{(i)}, M(X^{(i)})) + d(X^{(i)} \oplus X^{(i)}, T) \\ < \quad d(X, M(X)) + d(X \oplus X, T). \end{aligned} \qquad (4)$$

Any genome $X^{(i)}$ that minimizes the left hand side of inequality (3) or, better, inequality (4), over all genomes $X \in \mathbf{S}'$, and all trajectories between X and $M(X)$ (or $M(X^{(i)})$), is then a good initialization for a local hill-climbing search for an A, or for a pair $[A, M(A)]$, giving a local minimum for D. The details of the search vary from one empirical problem to another, but in our experience, there is often no better local minimum A than $X^{(i)}$ itself. If there is no such $X^{(i)}, i \geq 1$, then any $X \in \mathbf{S}'$ minimizes D.

4 A SMALL DATA SET ON MAIZE

It is generally agreed that the maize (*Zea mays*) genome underwent a genome doubling event some 11–16 million years ago (Gaut and Doebley, 1997). While some duplicated regions clearly attest to this event, there is no consensus on the exact inventory of such regions. Here we apply our method to infer the ancestor of the maize genome, with the rice (*Oryza sativa*) and sorghum (*Sorghum bicolor*) genomes as the two related diploids. For this purpose, we are concerned only with duplicated blocks in maize, and their single-copy counterparts in rice and sorghum, as extracted from the Gramene database (Jaiswal *et al.*, 2006), and not the remainder of each of the genomes.

In a previous study (Zheng *et al.*, 2006), we used Gramene to identify 34 syntenic blocks with two copies in maize and one copy each in sorghum and rice, though the partial nature of the maize genome sequence and the relative absence of sorghum sequence means that this genetic marker-based construction must be considered preliminary.

The genome halving algorithm usually involves some arbitrary choices in constructing the optimal ancestral tetraploid. In the case of the maize genome, this leads to more than 1 500 000 distinct solutions in **S**. The original data set not being very large (34 blocks in two genomes, 68 in maize), this exemplifies the extreme lack of uniqueness in the results of genome halving.

When we bring the diploid genomes to bear using Equation (2), however, testing all 1 500 000 elements of **S**, the set **S**′ contains only nine solutions. Thus there is a massive reduction of non-uniqueness induced by carrying out de-ploidization in phylogenetic context.

Searching for A and $M(A)$ along a trajectory from **S**′ towards the median using the criterion in inequality (4) led directly to the solution in Figure 3, which is not improved by local searching. Other trajectories from **S**′ towards the median gave three other solutions, with almost identical component distances. And other search methods (along trajectories to R_1 or R_2) provided a fifth solution, at a much greater distance, $d(T, A \oplus A) = 32$, from T.

For the schema in Figure 3, the given and inferred genomes, with synteny blocks evident, are depicted in Figure 4.

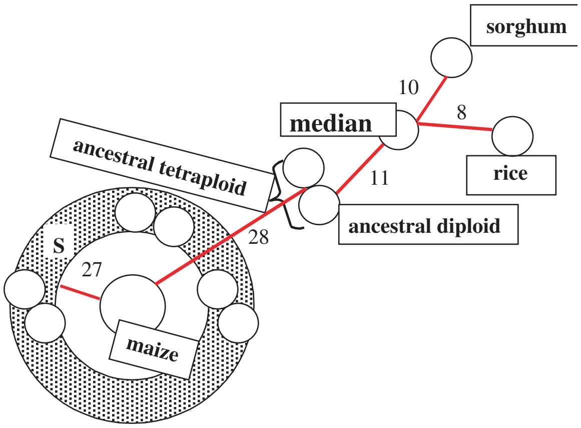

Fig. 3. Solution for the maize data.

5 TETRAPLOIDIZATION OF YEAST

Wolfe and Shields (1997) convincingly demonstrated an ancient tetraploidization of *Saccharomyces cerevisiae* a decade ago. Recently, the post-tetraploidization evolution of *S.cerevisiae* has been studied by comparison to the diploid genomes of *Ashbya gossypii* (Dietrich *et al.*, 2004) and of *Kluyveromyces waltii* (Kellis *et al.*, 2004), though without recourse to genome rearrangement or genome halving algorithms.

Each of these studies located a set of synteny blocks on the diploid genome, each block homologous to a pair of duplicate synteny blocks on the *S.cerevisiae* genome. These blocks were explicitly listed in the case of *K.waltii*, for which we could confirm 239 blocks, but only portrayed diagrammatically in the case of *A.gossypii*. We developed a protocol to tabulate the *A.gossypii* blocks based on this visual information, and obtained 409 blocks.

We then established a second protocol to align the blocks on *S.cerevisiae* corresponding to *K.waltii* blocks and those corresponding to *A.gossypii* blocks, sometimes dividing a long block from one diploid into shorter blocks corresponding to the other, and allowing ±2 extra ORFs on a block without throwing a correspondence into doubt. This protocol produced 263 blocks in both *K.waltii* and *A. gossypii*, each corresponding to a pair of duplicate blocks in *S.cerevisiae*.

Applying our method to this large data set produced the solutions in Figure 5. Because the time required for the median heuristic increases drastically with m, where we could handle 1.5×10^6 runs with $m = 34$ in the case of maize, we could only sample 2506 elements from **S** with $m = 263$, and found an **S**′ with only one element. To compensate for the sketchy coverage of **S**, we also examined several solutions of the genome halving algorithm where D was slightly suboptimal. Furthermore, we used the criterion in inequality (3) instead of the computationally more costly inequality (4) to locate A. Of interest is that one of the solutions has $A \in \mathbf{S}'$, though this was not one of the sampled genomes, but was found in the trajectory between a suboptimal solution B and $M(B)$.

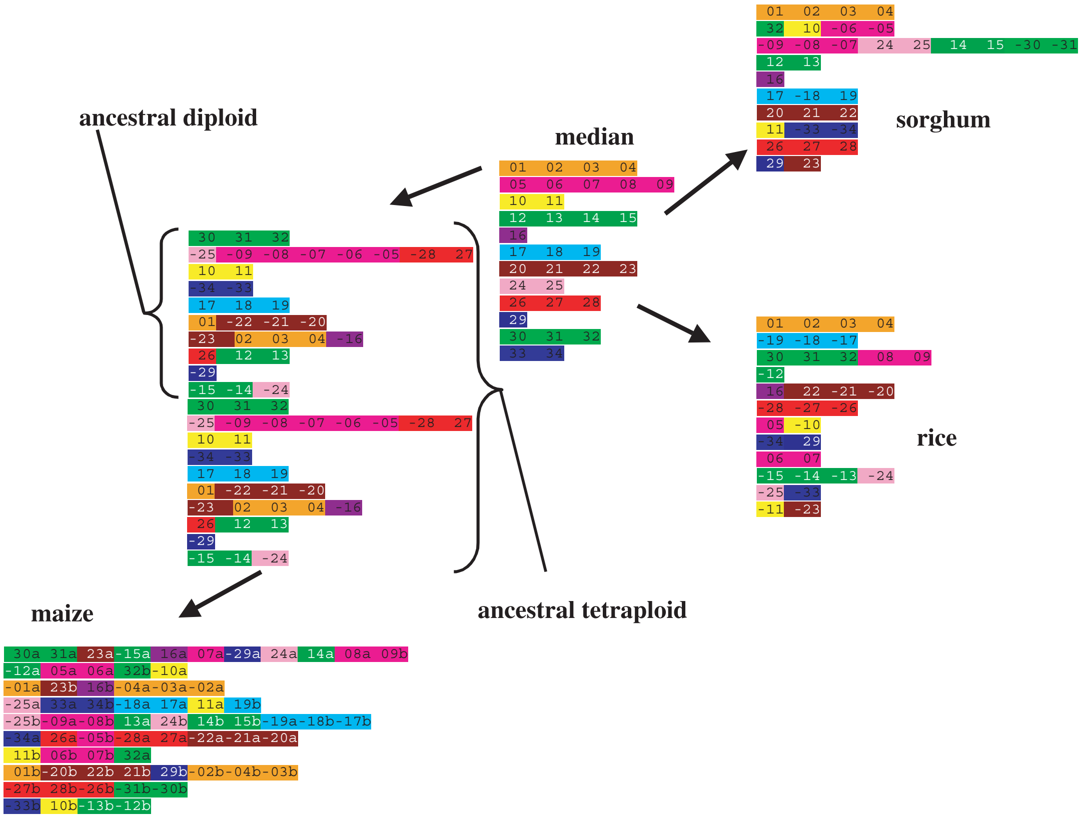

Fig. 4. Given and inferred cereal karyotypes and synteny blocks, color-keyed to the median genome.

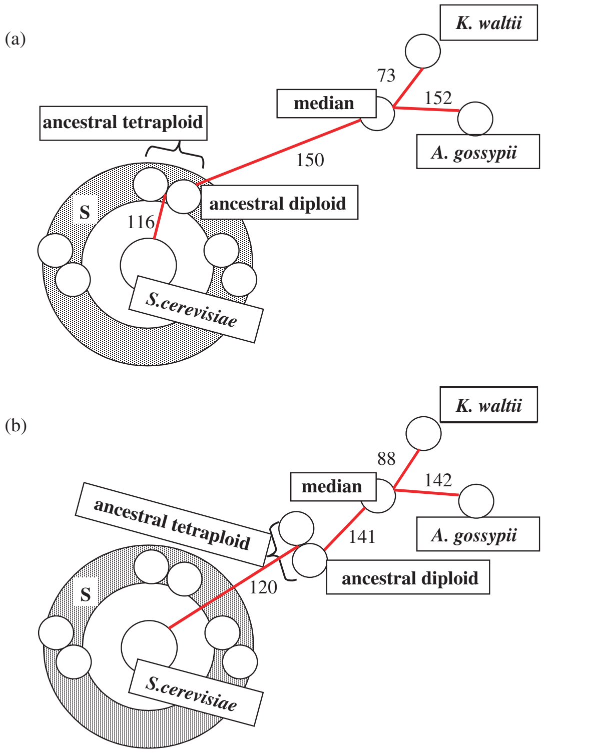

Fig. 5. Two solutions for the yeast data. (**a**) Solution ∈ **S**. (**b**) Solution ∉ **S** and detailed in Figure 7.

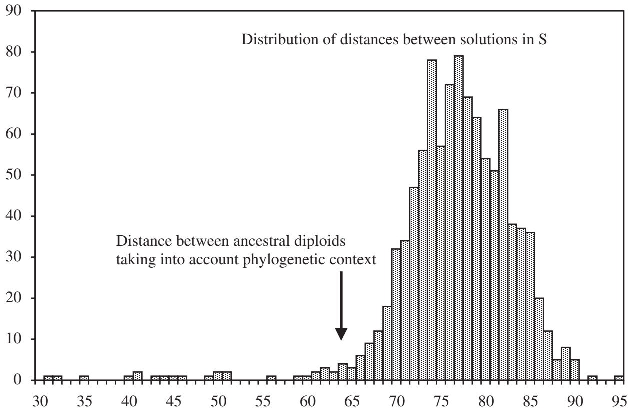

Fig. 6. Distribution of distances between genomes in **S**.

How different are these two solutions, summarized in Figure 5? If we calculate the rearrangement distance between them and compare it with randomly chosen pairs of genomes in **S** as in Figure 6, we see the distance between the two solutions is significantly smaller, although it is still large. Of course, it is possible that there is a unique, better, global optimum, but the impression gained from this example is that the present-day genomes do not contain very precise information on the position of the ancestral median in the space of genomes.

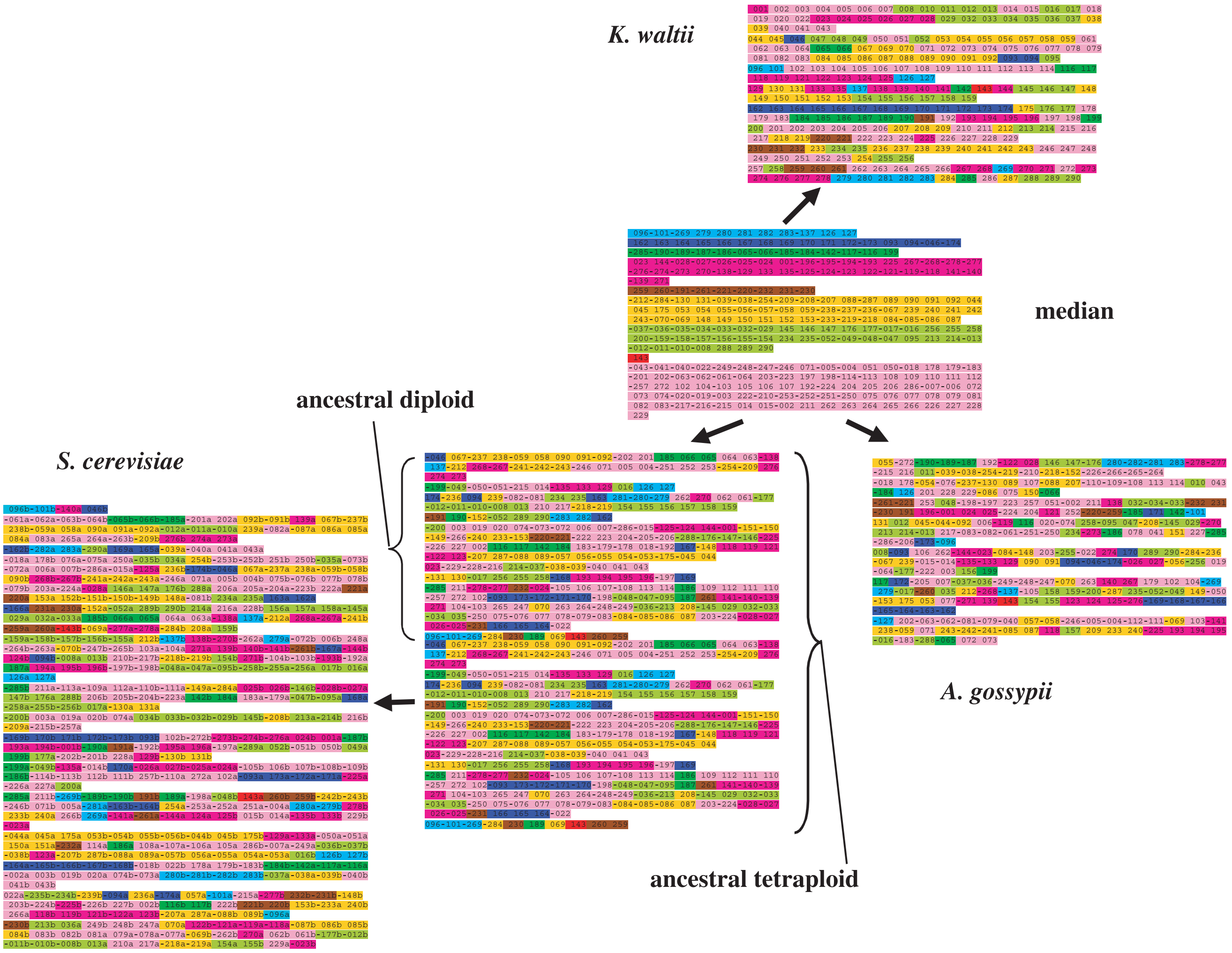

Fig. 7. Given and inferred yeast karyotypes and synteny blocks, color-keyed to the median genome. Long chromosomes are wrapped and chromosomes are separated by thin white space.

6 CONCLUSION

Among orthology assignment problems, the case of tetraploidy (and autopolyploidy in general) is rather unique in that DNA sequence information cannot help in partitioning the duplicate blocks into two sets, one from one copy of the original diploid, and the other set from the identical second copy, precisely because they were identical. This is not always the case with allopolyploidy since paralogs coming from one contributing polyploid would be more similar in DNA sequence amongst themselves than to paralogs from the other contributing polyploid. Thus our methodology could be made more precise in such cases by incorporating DNA sequence evidence insofar as allopolyploidy is concerned, but not autopolyploidy.

As mentioned in Section 3, there are many open problems to be solved before a general solution, even a heuristic one, is feasible for our simple model of polyploidy. And there are many more problems for a general model allowing for autopolyploidy by means other than tetraploidization.

Algorithmically, a difficult problem would be to replace our sequential procedure by a single algorithm searching for the pair $[A, M]$ that minimizes $D(T, R_1, R_2)$. This would be a hard problem, given that the median problem itself is NP-hard. Modifying the halving algorithm so that it could take account of both R_1 and R_2, while retaining optimality of the ancestral diploid, might be a good strategy for avoiding the construction of the entire set **S**, but would not mitigate the complexity of the median step.

ACKNOWLEDGEMENTS

Research supported in part by a grant to D.S. from the Natural Sciences and Engineering Research Council of Canada (NSERC). D.S. holds the Canada Research Chair in Mathematical Genomics and is a Fellow of the Evolutionary Biology Program of the Canadian Institute for Advanced Research.

Conflict of Interest: none declared.

REFERENCES

Bergeron,A. *et al*. (2006) A unifying view of genome rearrangements. In Bücher,P. and Moret,B.M.E. (eds.), *Algorithms in Bioinformatics. Proceedings of WABI 2006*. *Lect. Notes Comput. Sci.*, **4175**, 163–173.

Bourque,G. and Pevzner,P. (2002) Genome-scale evolution: reconstructing gene orders in the ancestral species. *Genome Res.*, **12**, 26–36.

Dietrich,F.S. *et al*. (2004) The *Ashbya gossypii* genome as a tool for mapping the ancient *Saccharomyces cerevisiae* genome. *Science*, **304**, 304–307.

El-Mabrouk,N. and Sankoff,D. (2003) The reconstruction of doubled genomes. *SIAM J. Comput.*, **32**, 754–792.

Gaut,B.S. and Doebley,J.F. (1997) DNA sequence evidence for the segmental allotetraploid origin of maize. *Proc. Natl Acad. Sci. USA*, **94**, 6809–6814.

Jaiswal,P. *et al*. (2006) Gramene: a bird's eye view of cereal genomes. *Nucleic Acids Res.*, **34**, D717–D723. URL: http://www.gramene.org

Kellis,M. *et al* (2004) Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae*. *Nature*, **428**, 617–624.

Wolfe,K.H. and Shields,D.C. (1997) Molecular evidence for an ancient duplication of the entire yeast genome. *Nature*, **387**, 708–713.

Zheng,C. *et al*. (2006) Genome halving with an outgroup. *Evol. Bioinformatics*, **2**, 319–326.