uOttawa

L'Université canadienne
Canada's university

FACULTÉ DES ÉTUDES SUPÉRIEURES ET POSTDOCTORALES

uOttawa

L'Université canadienne
Canada's university

FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES

Chunfang Zheng
AUTEUR DE LA THÈSE / AUTHOR OF THESIS

Ph.D. (Biology)
GRADE / DEGREE

Department of Biology
FACULTÉ, ÉCOLE, DÉPARTEMENT / FACULTY, SCHOOL, DEPARTMENT

Polynoids, Genome Halving and Rearrangement Phylogeny
TITRE DE LA THÈSE / TITLE OF THESIS

D. Sankoff
DIRECTEUR (DIRECTRICE) DE LA THÈSE / THESIS SUPERVISOR

CO-DIRECTEUR (CO-DIRECTRICE) DE LA THÈSE / THESIS CO-SUPERVISOR

EXAMINATEURS (EXAMINATRICES) DE LA THÈSE / THESIS EXAMINERS

G. Drouin

M. Turcott

E. Kranakis

L. Zhang

Gary W. Slater
Le Doyen de la Faculté des études supérieures et postdoctorales / Dean of the Faculty of Graduate and Postdoctoral Studies

POLYPLOIDS, GENOME HALVING AND REARRANGEMENT PHYLOGENY

Chunfang Zheng

Thesis Submitted to the Faculty of Graduate and Postdoctoral Studies

In partial fulfillment of the requirements

for the degree of

Doctor of Philosophy in Biology[1]

Department of Biology

Faculty of Science

University of Ottawa

[1]The Ph.D. Program is a joint program with Carleton University, administered by the Ottawa-Carleton Institute of Biology

Your file Votre référence
ISBN: 978-0-494-59481-0
Our file Notre référence
ISBN: 978-0-494-59481-0

Abstract

The basic rearrangement phylogeny methods require that the genomic content be the same in all the organisms being compared, and so are not applicable when one or more of the genomes being compared derive from ancestral whole genome doubling (WGD) events. In this thesis I developed algorithms for rearrangement phylogeny for sets of related genomes that include both descendants of WGD and unduplicated genomes. Furthermore I investigated the properties of these algorithms and validated them by applying them to real data.

I defined 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.

The genome halving algorithm of El-Mabrouk and Sankoff efficiently reconstructs an ancestral pre-doubling genome from the chromosomal distribution of duplicate genes created by this event and remaining today in the descendant genome. However, this algorithm can produce many alternate optimal solutions. To reduce this non-uniqeness, hopefully to only one solution, I developed the guided genome halving algorithm. This rapidly and accurately constructs an optimal ancestor closest to one or more outgroups.

As I refined this algorithm, I applied it to successively larger data sets, increasing in size over more than two orders of magnitude. Thus I constructed the ancestors

of cereals, based on duplicate markers in maize and using rice and sorghum as outliers. I reconstructed genomes corresponding to the ancestral nodes of yeasts in the Saccharomyces complex, as well as the ancestor of poplar, based on grapevine and papaya as outgroups.

I studied two cases involving two WGD descendants, one where the doubling precedes a speciation event and another where doubling occurs independently in both lineages initiated by a speciation event. I developed combinatorial algorithms permitting us to decide which of these options best explains the data.

To take into account the massive loss of genes following ancestral genome doubling, I developed a method to incorporate the defective gene sets into consideration and evaluate the effects of these lost genes on the reconstruction of the ancestor.

Résumé

Les méthodes phylogénétiques basées sur l'ordre de gènes, dont la *phylogénie de réarrangement*, exigent que les espèces comparées ont toutes le même contenu génomique. Ces méthodes ne sont donc pas applicables lorsque l'un ou plusieurs des génomes ont des origines dans des événements de tétraploïdisation (ou duplication du génome entier – DGE). Dans cette thèse, j'ai développé des algorithmes pour la phylogénie de réarrangement applicables aux ensembles de génomes qui comprennent à la fois des descendants de DGE ainsi que des génomes non-dupliqués. De plus, j'ai étudié les propriétés de ces algorithmes et je les ai vérifié en les appliquant à de véritables données.

J'ai repéré plusieurs configurations locales de génomes dupliqués et non-dupliqués dans une phylogénie, dont chacun exige une stratégie spécifique qui intègre les algorithmes de distance génomique, de de-tétraploïdisation et de médiane génomique.

L'algorithme de de-tétraploïdisation d'El-Mabrouk et Sankoff reconstruit de façon efficace les chromosomes de l'ancêtre du tétraploïde à partir de la distribution actuelle de gènes dupliqués qui restent aujourd'hui dans le génome du descendant. Toutefois, cet algorithme peut produire de nombreures solutions optimales très différentes. Pour réduire cette non-unicité, j'ai développé un algorithme de *de-tétraploïdisation guidée*. Ceci construit rapidement et avec précision à un ancêtre le plus proche de l'un ou plusieurs génomes diploïdes de référence.

J'ai du raffiner cet algorithme à force de l'appliquer successivement à des plus grands ensembles de données, à travers de plus de deux ordres de grandeur. Ainsi, j'ai construit les ancêtres des céréales, sur la base de marqueurs dupliqués en maïs et présents aussi dans le riz et le sorgho comme génomes de référence. J'ai reconstruit génomes ancestraux des levures dans le groupe de Saccharomyces, ainsi que l'ancêtre

de peuplier, avec la vigne et la papaye comme génomes de référence.

J'ai étudié deux cas impliquant deux DGE descendants, l'un où la DGE précède la spéciation et l'autre où une duplication se produit de façon indépendante dans les deux lignées initiées par un événement de spéciation. J'ai élaboré des algorithmes combinatoires permettant de décider laquelle de ces options mieux explique les données.

Pour tenir compte de la perte massive de gènes ancestraux conséquent à un DGE, j'ai proposé une méthode permettant d'intégrer les ensembles incomplèts d'orthologuls dans l'analyse et d'évaluer les effets de la perte des gènes manquants sur la reconstruction de l'ancêtre.

Acknowledgements

First, I sincerely thank my advisor Dr. David Sankoff, without whom this work would not have been possible. I am grateful to him for providing me the opportunity to explore and discover in the new area of genome rearrangement. I very much appreciate his guidance, continuous encouragement and financial support over the years from my Masters program to my Ph.D. Moreover, I extend thanks to all the members of the Sankoff lab: Qian Zhu, Zaky Adam, Robert Warren, Wei Xu, Zhenyu Yang, Adriana Muñoz, Shengang Li, Ximing Xu and Ghada Badr. I have benefited much from their collaboration, friendship and helpful discussions. To these I add external collaborators Dr. Vicky Choi, Dr. Eric Tannier, Dr. Victor Albert, Dr. P. Kerr Wall and Dr. Ken Wolfe.

Special thanks go to the Department of Biology of the University of Ottawa for giving me the freedom to specialize in bioinformatics and to focus on this area, and to my committee members: Dr. Guy Drouin, Dr. Pat Morin of Carleton University and Dr. Linda Bonen, for their helpful suggestions and for being available to help me. I am grateful to Dr. Marcel Turcotte of the School of Information Technology and Engineering for introducing me to the field of bioinformatics and for valuable help and advice. And thank you to Ginette Soutar and Doreen Smith, for administrative support and assistance during my graduate studies.

I am grateful to my husband Hong Liu for his trust, support and patience

throughout this work and in my daily life.

My research was supported financially by Natural Sciences and Engineering Research Council of Canada (NSERC), Canada Graduate Scholarships (OGS) and grants awarded to Dr. David Sankoff by NSERC, Le Fonds québécois de la recherche sur la nature et les technologies (FQRNT), the Canada Research Chair program and the University of Ottawa.

Dedication

To my daughter, Karen, and my son, Alex, for giving me dreams and courage.

Contents

List of Tables

List of Figures

Chapter 1

General introduction

1.1 Rearrangement phylogeny

The entire complement of genetic material carried by an individual is called the genome. With the exception of some viruses, each genome is physically composed of molecules of *deoxyribonucleic acid* (DNA). In higher organisms, the nuclear genome is partitioned among a number of chromosomes, each chromosome carrying its own characteristic DNA. The DNA consists of two complementary strands twisted around each other to form a right-handed double helix. A gene is a segment of DNA sequence with a specific function such as coding for a protein or RNA structure. Genes can be ordered according to the location of their DNAs on a chromosome. Because genes may be located on either strand of a DNA molecule, a sign (+/-) is usually used to indicate on which strand a gene is found. Gene content and gene order are generally identical in all the cells of an individual organism and, for our purposes, in all the individuals of a species.

If two genes are similar due to a shared common ancestor gene, they are called homologs. There are two kinds of homologs, namely orthologs and paralogs. Or-

thologs are genes in different species that evolved from a common ancestral gene by speciation. Paralogs are genes in a single genome related by a duplication event in some ancestor of that genome. Single orthologs tend to retain the same function over long periods of time in the course of evolution, whereas some paralogs evolve new functions relatively quickly, often related to the original one, through the processes of neofunctionalization or subfunctionalization.

Since the genome is structurally specific to each species, insofar as it changes slowly over time, genome comparison among related species can provide us with much evidence about evolution. This includes reconstructing the phylogenetic tree and the changes that have occurred over the branches of this tree.

A phylogenetic tree represents the evolutionary relationships among a set of related species. Reconstructing the phylogenetic tree is a fundamental problem in computational biology. The genomes of two closely related species are more similar than those of two distantly related species. Consequently, by comparing the genomic similarities among species, we can deduce the evolutionary history of these species in terms of speciation and divergence processes.

We may distinguish two types of genomic change. One is local, such as point mutations or small insertions and deletions in the DNA sequence. The other is genome rearrangement, involving the movement of large or small fragments of the chromosome, such as inversion or transposition. There are correspondingly two approaches to comparing genomes when reconstructing the phylogenetic tree. The traditional one is to count local changes such as point mutations. The other way, which is adopted in this thesis, is to compare the order of the genes or other markers in genomes, as they have been affected by genome rearrangement.

Genome rearrangements are an important aspect of speciation and the evolutionary divergence of species. Because of genome rearrangement, the gene order of

different species can be quite different while the gene content is almost identical. Although genome rearrangements are much less frequent than point mutation, they are important because they involve major segments of the genome and are likely to be involved in speciation during evolution. Rearrangement phylogenetics constructs phylogenetic trees by comparing the gene order data of different species. It provides an approach to phylogenetics independent of that based on nucleotide or amino acid sequence divergence.

The basic rearrangement phylogeny methods [26, 139, 170] require that the genomic content be the same in all the organisms being compared, so that every marker in one genome can be identified with a single orthologous counterpart in each of the others, though adjustments can be made for a limited amount of deletion, insertion [205] and duplication [171] of chromosomal material. This thesis is largely devoted to overcoming the identical-content limitation in important contexts where it is completely inappropriate.

1.2 Whole genome duplication

Polyploidization is a special class of mutation that multiplies the content of the genome. It can occur via several mechanisms, such as genomic doubling, gametic non-reduction and polyspermy. Polyploidy is widespread in plants (e.g., 30-80% of angiosperms are descended from polyploids) and animals [147]. It plays an important role in ecology and evolution, allowing populations of species to occupy new niches and to develop new genetic networks, and often leads to speciation, with the polyploid becoming reproductively isolated from the original diploid. The most frequent polyploidy event is tetraploidization, after which meiosis is characterized not by the normal pairings of one maternal and one paternal chromosome, but by quadriva-

lent alignment of chromosomes or other combinations. Tetraploidization is followed by a period of re-diploidization, where distinct chromosomal 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 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.

In this thesis, the terms "whole genome duplication" and "tetraploidization" are used largely interchangeably, even though the latter connotes more specific mechanisms than the former. Indeed, some authors eschew the term "tetraploidization" in order to avoid implying the process by which the genome became doubled, or to avoid the anomaly of applying the label "tetraploid" to some generally haploid organism, like some yeasts, that has undergone WGD. Since my mathematical and algorithmic approach in no way depends on these considerations, I hope the reader will forgive these usages.

Genomic distance or rearrangement phylogeny algorithms are not applicable when some of the species descend from a tetraploid, since there may be many two-to-one relationships between markers in the doubled genomes and those in the unduplicated species, whose divergence predates the doubling event. As we stressed above, these algorithms require one-to-one correspondences.

However, reconstruction of ancestral genomes can potentially be achieved through comparative analysis of genome content and structure for extant species, combining information from gene duplication histories (developed through analysis of sequence similarities or gene trees) and chromosomal gene locations to identify duplicated blocks of genes or DNA sequence dispersed among the chromosomes in one or more

species. Note that sequence analysis tools are not directly applicable to this problem because all pairs of duplicates in doubled genomes were generated at the same historical moment.

The genome halving problem is thus to reconstruct the ancestral genome on the basis of a decomposition of the present-day genome into a set of apparently duplicated blocks of genes or DNA sequence dispersed among the chromosomes. A quantitative approach to this problem was first discussed by Seoighe and Wolfe [180] in the context of the genome doubling of the ancestor of the yeast *Saccharomyces cerevisiae*. At the same time, motivated by studies of genome duplication in early vertebrates [143], El-Mabrouk and colleagues [47, 48, 50] published a series of papers on the combinatorial optimization approach to the problem, culminating in a general solution [51]. Further refinements have been published by Alekseyev and Pevzner [6], Warren and Sankoff [219] and Mixtacki [136].

Seoighe and Wolfe [180] noted an extreme degree of non-uniqueness associated with the solution to the genome halving problem. Many, rather different, genomes serve equally well as the pre-doubling ancestor, according to the optimization criterion. These authors suggested that this difficulty could be attenuated through the use of a reference genome, or outgroup. The suggestion to use a reference genome was taken up to study the post-tetraploidization evolution of *S. cerevisiae*, both in reference to the genome of *Ashbya gossypii* [44] and to that of *Kluyveromyces waltii* [86], though without recourse to genome rearrangement or genome halving algorithms. Similar research compared mammalian genomes with the tetraploid ancestor of the puffer-fish *Tetraodon nigroviridis* [80].

Genome halving for the reconstruction of the pre-doubling ancestral genomes is important not only for phylogenetic purposes, but also to improve our ability to resolve correlated gene arrangements among taxa [27], which in turn provides a valu-

able framework for inference of shared ancestry of genes and for the utilization of findings from model organisms to study less well-understood systems. In addition it can also reveal patterns of differential gene loss following genome duplication, differential gene retention associated with evolution of some morphological complexity, and unexpectedly large variation among taxa in DNA substitution rates [207].

Complications of genome halving include not only the non-uniqueness of the solutions, but also the fact that whole-genome duplication may be followed, over evolutionary time, by genome downsizing (paralog extinction and the loss of noncoding DNA), sharply reducing the amount of data available to the algorithms, as well as continual gene duplication events that do not involve WGD.

1.3 The objectives of the thesis

My research aims to provide an overall strategy and to develop algorithms for rearrangement phylogeny for sets of related genomes that include some that have undergone doubling.

I pose the computational problems only within the framework of the "small" phylogenetic problem[1], 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. Then I apply the algorithms to reconstruct the ancestors of maize, of *Saccharomyces* and of *Populus*, each of which shows strong evidence of an ancient whole genome duplication [60, 213, 227].

[1]The "large" problem of constructing the branching order of the phylogeny is a separate and independent question. For the data sets I consider, there is little debate over the tree structure.

1.4 The data

1.4.1 The cereals

It is generally agreed that the maize genome underwent a genome doubling event some 11-16 million years ago [60]. While some duplicated regions clearly attest to this event, there is no consensus on the exact inventory of such regions. Moore *et al.* and Wilson *et al.* [137, 225] presented two largely consistent views of syntenic blocks across the cereals based on the mapping evidence at the time. These included 14 and 19 duplicated blocks in the maize genome. Gaut [61] gave a more comprehensive account of the pattern of 23 duplicated regions, based on maize genomic sequence data in 2001. He did not completely establish the relative position of all the syntenies on the chromosomes in this work.

Even now that the rice genome has been sequenced, and the maize genome project is well-advanced, it is no trivial matter to identify the duplicate blocks resulting from the tetraploidization event. The maize genome has many other duplicated segments dating from periods both after and before the tetraploidization and even before the divergence from the other cereals. This is complicated by post-tetraploidization genome rearrangement events, deletions and insertions of genetic material, transpositions of genes or larger segments from one site on the genome to another, and loss of homology between the parts of the duplicated regions.

The databank which has the most information on the syntenies among the cereal genomes is Gramene [82]. The current version at time of writing is release 21. From this we can obtain a conservative (i.e., confined to high homology regions only) estimates of duplicate blocks by comparison with the rice genome. For example, in Figure 1.1, we can visually identify large duplicated regions in chromosomes 1 and 9,

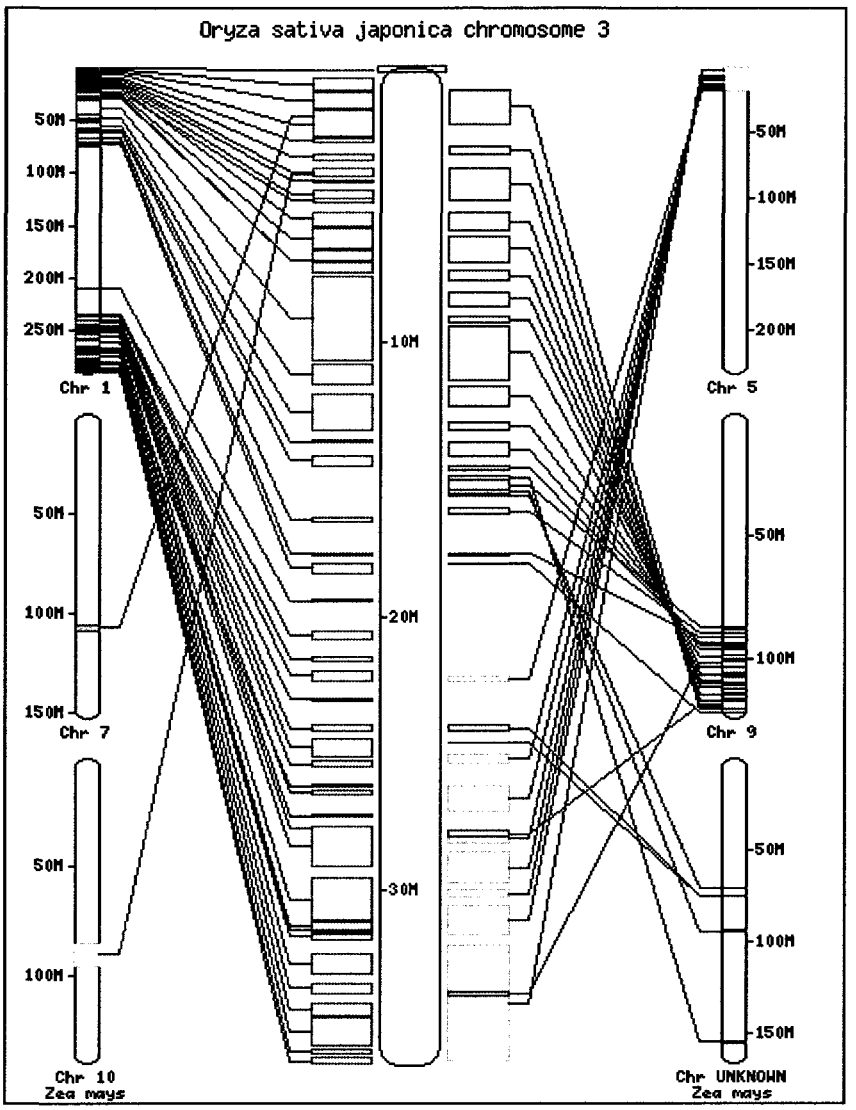

Figure 1.1: Syntenies between rice chromosome 1 and maize chromosomes, as produced by Gramene [82].

chromosomes 1 and 5, and possibly a number of smaller ones, all by virtue of their common homology with regions of rice chromosome 3.

Unfortunately, there is as yet no comparison of syntenic blocks between sorghum and the other genomes on Gramene. However, there are extensive mapping data of various kinds of markers. We bolstered our preceding data collection by searching sets

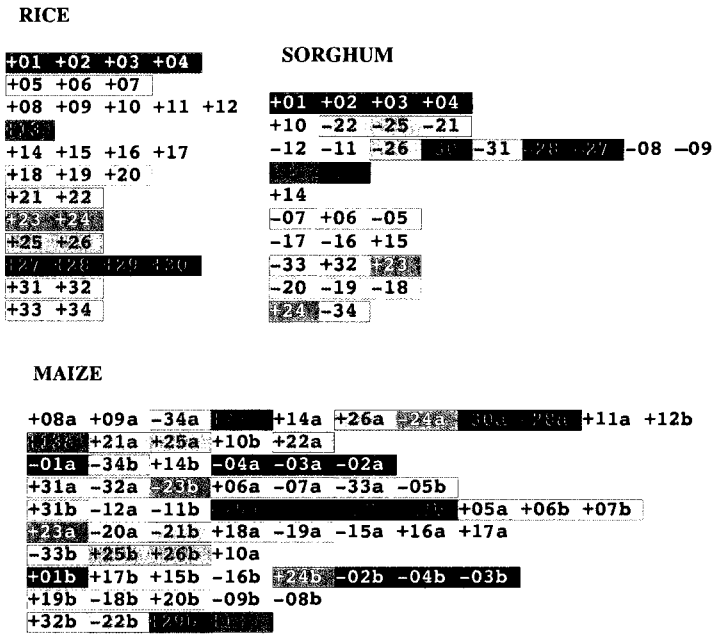

Figure 1.2: Order of syntenic blocks in rice, sorghum and, in two copies each, maize.

of duplicate markers in maize that had single copies in sorghum and rice, comparing mainly the Patterson 2003 genetic map for sorghum, the IBM2 Neighbours 2004 and Cornell Wilson 1999 genetic maps for maize and the Annotated Nipponbare Sequence 2006 sequence map for rice. All the markers satisfying these criteria fell into the rice-maize syntenies established previously. Based on these criteria, i.e., markers identified as homologous in Gramene, with a single copy in each of rice and sorghum and two copies in maize, plus the requirement that the maize and rice copies fall into the appropriate, previous identified, rice-maize syntenic blocks, we could now identify 34 syntenic blocks as basic data for our reconstruction. These data are depicted in Figure 1.2, but should be considered to constitute a working hypothesis; definitive data must await the finishing of the maize genome, the sequencing of the sorghum genome, and the further application of global alignment and synteny block construction methods.

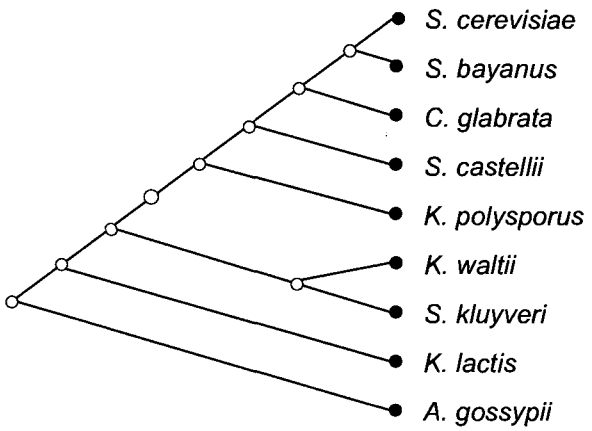

Figure 1.3: Phylogenetic tree of yeasts, provided by [31, 95, 177].

1.4.2 Yeast

In 1997, Wolfe and Shields [227] discovered an ancient genome doubling in the ancestry of *Saccharomyces cerevisiae*, after this organism became the first eukaryote to have its genome sequenced [62]. According to [31,95,177], five out of the nine yeasts in the phylogenetic tree of Fig 1.3 are doubling descendants which share the same doubled ancestor. The Yeast Gene Order Browser [31](YGOB) provides the gene order of these nine yeasts as well as a reconstruction of the doubled ancestor. Of particular interest is that YGOB contains inferences about which of the two duplicates in one genome is orthologous to which duplicate in the other genome. This will prove to be of great help in validating our methods.

1.4.3 Angiosperms

The poplar (*Populus trichocarpa*) has undergone whole genome duplication followed by extensive chromosomal rearrangement, and is one of four angiosperm genomes, along with those of *Carica papaya* (papaya), *Vitis vinifera* (grapevine) and *Arabidopsis thaliana*, whose sequences have been published to date, shown in Figure 1.4.

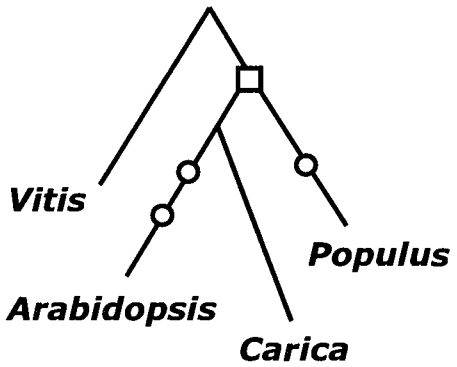

Figure 1.4: Phylogenetic relationships among angiosperms with sequenced genomes. The circles indicate likely whole genome doubling events. The circle in the *Populus* lineage, representing the locus of the WGD event at the origin of the willow-poplar family, and the square, representing the ancestor of the rosid dicotyledons, indicate the target ancestors we reconstruct in this paper.

Annotations for the *Populus*, *Vitis* and *Carica* genomes were obtained from databases maintained by the U.S. Department of Energy's Joint Genome Institute [213], the French National Sequencing Center, Genoscope [81], and the University of Hawaii [135], respectively. An all-against-all BLASTP search was run on a data set including all *Populus* and *Vitis* protein coding genes, and orthoMCL [113] was used to construct 2104 full and 4040 defective gene sets, in the first case, denoted PPV, containing two poplar paralogs (genome P) and one grape ortholog (genome V), and in the second case, denoted PV or PP, missing a copy from either P or V. This was repeated with *Populus* and *Carica*, genomes P and C, respectively, to obtain 2590 full (PPC) and 4632 defective (PC or PP) sets. The location on chromosomes (or contigs in the case of *Carica*) and orientation of these paralogs and orthologs was used to construct our database of gene orders for these genomes. Contigs containing only a single gene were discarded from the *Carica* data.

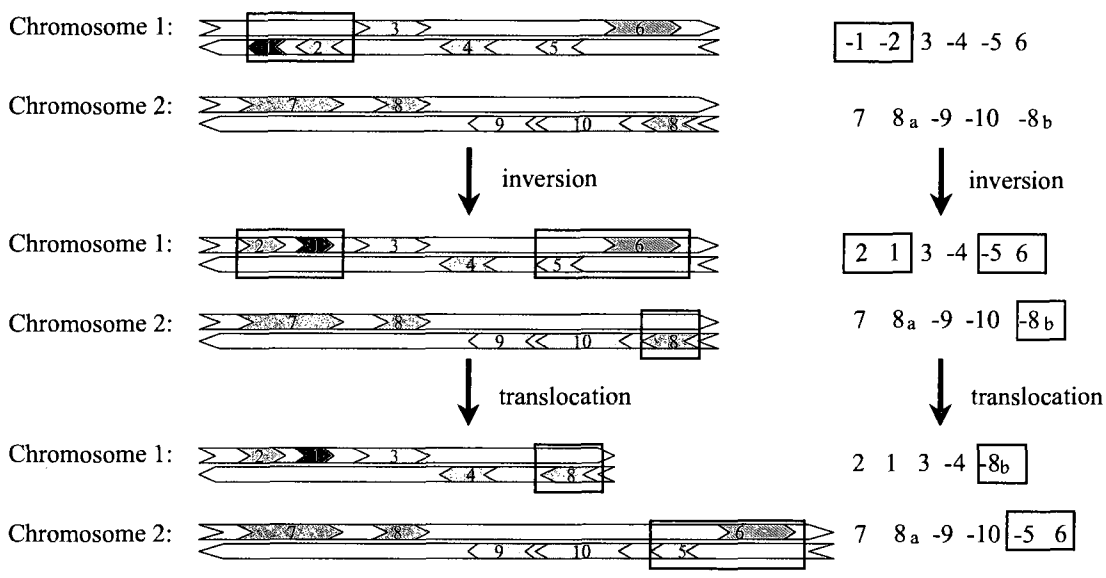

Figure 1.5: Genomes and genome rearrangement operations

1.5 Existing algorithmic resources

Given a set of genomes which include the doubling descendants, my research objectives include the development of effective methods for reconstructing the gene order of the ancestral genomes and thus infer the cost of the phylogeny in terms of rearrangement events. Here I will discuss some existing basic elements of the solution, i.e. genomic distance, genome halving and genome median. These existing resources are also the prerequisite for understanding the research work in this thesis.

1.5.1 Genomes and rearrangement operations

A genome can be modeled as a set of strings, each string representing one chromosome, by writing the genes in the order they appear in this chromosome. The sign "+/-" before a gene indicates on which strand this gene is located. Subscripts ("a" and "b") are used to differentiate two copies of paralogs in a genome if this genome contains

duplicate genes.

The biologically-motivated operations generally include

- inversions (implying as well change of sign, i.e., change of strand) of chromosomal segments, e.g.,

  $h_1 \cdots h_u \cdots h_v \cdots h_m \rightarrow h_1 \cdots -h_v \cdots -h_u \cdots h_m,$

- reciprocal translocations, e.g.,

  $h_1 \cdots h_u \cdots h_l, k_1 \cdots k_v \cdots k_m \rightarrow h_1 \cdots k_v \cdots k_m, k_1 \cdots h_u \cdots h_l,$

- chromosome fission e.g.,

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

- chromosome fusion e.g.,

  $h_1 \cdots h_l, k_1 \cdots k_m \rightarrow h_1 \cdots h_l k_1 \cdots k_m$

1.5.2 Genomic distance

Genome distance is used to measure gene order differences between two related genomes. It is defined between two species of the same ploidy, where one-to-one orthology relations have been identified between the genes in the two genomes. The well studied genome distances are *breakpoint distance* and *rearrangement distance.* To mathematically handle the sign of genes, each gene A is replaced by its two *extremities*, its tail (A_t) and its head (A_h). If the sign of A is "+", then the tail extremity precedes the head extremity. Otherwise if the sign of A is "-", then A_h precedes A_t. An *adjacency* is a pair of unordered extremities that contains the information on which extremities of two neighbouring genes are next to each other. For example, if G and H are two successive genes on a chromosome, and g and h are their two extremities next to each other, irrespective of whether they are heads or tails,

then $\{g, h\}$ (or equivalently $\{h, g\}$) constitutes an adjacency. There are two special extremities on a linear chromosome, called *telomeres*, namely the initial extremity of the first gene (i.e., the one that is not adjacent to any other gene) and similarly the final extremity of the last gene.

Breakpoint distance

When two genomes Π and Γ on a set of n genes are compared, an adjacency or telomere is *in common* if it appears in both genomes, otherwise it is a breakpoint. The breakpoint distance $d_B(\Pi, \Gamma)$ indicates how many breakpoints are in genome Π when comparing it to genome Γ. Let $a(\Pi, \Gamma)$ be the number of common adjacencies, $e(\Pi, \Gamma)$ be the number of common telomeres of Π and Γ, then [208]

$$d_{BP}(\Pi, \Gamma) = n - a(\Pi, \Gamma) - \frac{e(\Pi, \Gamma)}{2}. \tag{1.5.1}$$

Note that

$$d_{BP}(\Pi, \Gamma) = d_{BP}(\Gamma, \Pi). \tag{1.5.2}$$

Rearrangement distance

The rearrangement distance is the smallest number of rearrangement operations necessary to transform one genome into another. As described in [231], the rearrangement distance can be calculated in linear-time by means of the bicoloured breakpoint graph. To construct the breakpoint graph of genomes Π with h chromosomes and Γ with k chromosomes on a set of n genes, we first convert each telomere g in genome Π into an adjacency of form $\{g, \text{capA}\}$, where "capA" is an artificial extremity, which must then be subscripted "$2i - 1$" or "$2i$" if it is on the i-th chromosome. We do the same thing to the telomeres in genome Γ, by including them in adjacencies with "capBs".

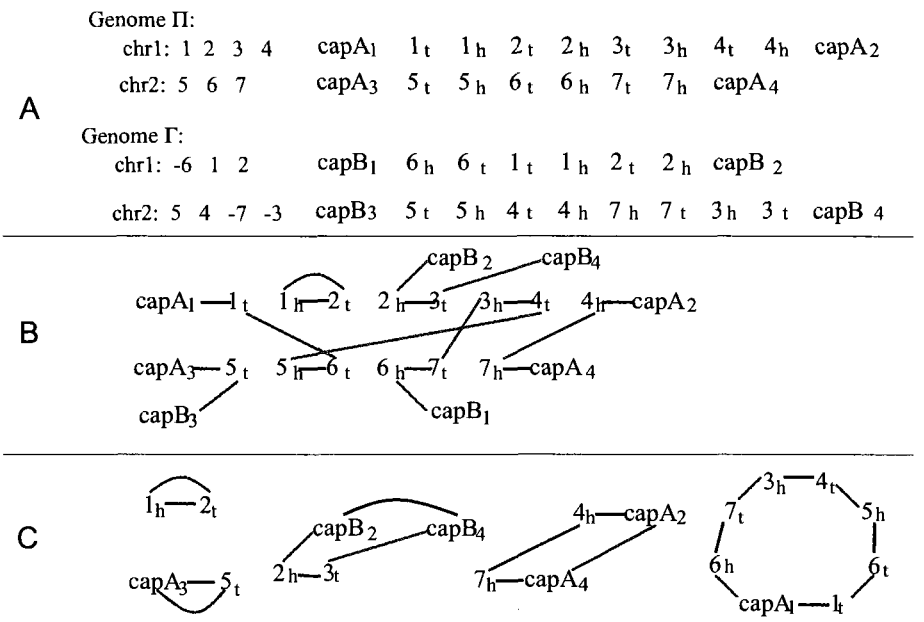

Figure 1.6: Breakpoint graph. A: Genomes Π and Γ. B: Vertices and edges of breakpoint graph. C: Convert paths into cycles.

Then the vertices of the breakpoint graph include the extremities in Π (or Γ) and capAs and capBs. The edges of the breakpoint graph are determined by the adjacencies in genomes Π and Γ. For each adjacency in genome Π, we add a black edge in the breakpoint graph to connect the two vertices that make up this adjacency. And we add gray edges to the breakpoint graph according to the adjacencies in genome Γ.

Each vertex is then incident to exactly one black and one gray edge, except for the capAs and capBs (each capA/capB is only incident to one black/gray edge). The bicoloured graph decomposes uniquely into a number of alternating colour cycles plus $h+k$ alternating colour paths terminating in cap vertices. We convert each path into a cycle according to the following rules: if both of the terminals are capAs, i.e. $capA_i$ and $capA_j$, add a gray edge to connect $capA_i$ and $capA_j$; if both of the terminals are capBs, i.e.$capB_i$ and $capB_j$, add a black edge to connect $capB_i$ and $capB_j$; if one terminal is capA and another is capB, i.e. $capA_i$ and $capB_j$, then remove the

$capB_j$ and close the path by the gray edge that was previously incident to $capB_j$. Figure 1.6 shows a breakpoint graph and its paths and cycles. Let $c(\Pi, \Gamma)$ be the number of cycles in the breakpoint graph. Yancopoulos *et al.* [231] proved that the rearrangement distance

$$d_R(\Pi, \Gamma) = n - c(\Pi, \Gamma) \tag{1.5.3}$$

1.5.3 Genome halving.

Let T be a genome consisting of ψ chromosomes and $2n$ genes $g_a^1 \cdots, g_a^n; g_b^1, \cdots, g_b^n$, dispersed in any order on the chromosomes. For each i, we call g_a^i and g_b^i "duplicates", but there is no particular property distinguishing all elements of the set of g_a^i in common from all those in the set of g_b^i. A potential "doubled ancestor" of T is written $A' \oplus A''$, and consists of 2χ chromosomes, where some half (χ) of the chromosomes, symbolized by the A', contains exactly one of g_a^i or g_b^i for each $i = 1, \cdots, n$. The remaining χ chromosomes, symbolized by the A'', are each identical to one in the first half, in that where g_a^i appears on a chromosome in the A', g_b^i appears on the corresponding chromosome in A'', and where g_b^i appears in A', g_a^i appears in A''. We define A to be either of the two halves of $A' \oplus A''$, where the subscript a or b is suppressed from each g_a^i or g_b^i. These χ chromosomes, and the n genes they contain, $g^1, \cdots, g^n$ constitute a potential "doubled ancestor" of T.

The genome halving problem for T is to find an A for which some $d(A' \oplus A'', T)$ is minimal. Here I only describe the initial part of El-Mabrouk's linear-time algorithm for solving this problem [51]. This part suffices to calculate the minimal $d(A' \oplus A'', T)$. The complicated details of actually reconstructing the minimizing A can be found in the original publication. I use this in the early chapters of this thesis but abandon it

in favour of new heuristics to be described in the later chapters.

The genome halving algorithms [51] also makes use of the breakpoint graph, but the problem here is building the breakpoint graph where only vertices and black edges are known. The vertex set of this breakpoint graph includes the extremities of $2n$ genes and 2ψ caps O $(O_1, O_2, \cdots, O_{2\psi})$, where each cap forms an adjacency with some telomere. The black edges of this graph are the adjacencies of genome T. If ψ is odd, a "null" chromosome is added, which means there are two extra caps ($O_{2\psi+1}$ and $O_{2\psi+2}$) and one extra black edge, which connects $O_{2\psi+1}$ and $O_{2\psi+2}$ in the breakpoint graph.

Since subscripts "a" and "b" are used to differentiate two copies of paralogs, "a" is defined to be the *counterpart* of "b" and "b" is the counterpart of "a". A extremity X always has a counterpart extremity $\bar{X}$ because each gene in T has a duplicate. For duplicates g_a^i and g_b^i, if $X = g_{at}^i$, then $\bar{X} = g_{bt}^i$; and if $X = g_{bh}^i$, then $\bar{X} = g_{ah}^i$. Of course $\bar{\bar{X}} = X$.

The vertices and edges are partitioned into *natural graphs* according to the following principle: If an edge $\{X,Y\}$ belongs to a natural graph, then so does some edge of form $\{\bar{X},Z\}$ and some edge of form $\{\bar{Y},W\}$.

If a natural graph has an even number of edges, as illustrated on the left of Figure 1.7, it can be shown that in all optimal ancestral doubled genomes, the edges coloured gray, representing adjacent vertices in the ancestor (recall the definitions in Section 1.5.2), and incident to one of the vertices in this natural graph, necessarily have as their other endpoint another vertex within the same natural graph.

Gray edges are added in pairs, so that each edge in the reconstructed doubled ancestor is duplicated, in accordance with the required output of the algorithm to produce two copies of the same ancestral genome. (Whether a is connected to b and $\bar{a}$ to $\bar{b}$ or a to $\bar{b}$ and $\bar{a}$ to b is immaterial, since the two versions of the same gene were

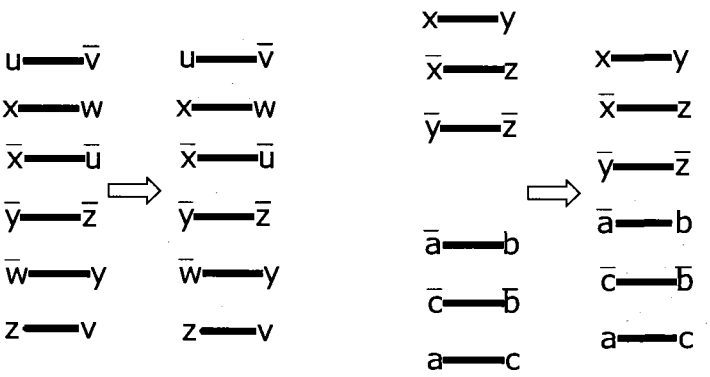

Figure 1.7: (left) Even-size natural graph completed by adding three pairs of gray edges in a way that maximizes the number of cycles. (right) Two odd-size natural graphs, containing x, y, z vertices and a, b, c vertices, respectively, combined into one supernatural graph so that three pairs of gray edges may be added.

originally identical.) These edges are added in such a way as to maximize the number of (alternating coloured) cycles and paths that make up this subgraph, in accordance with the definition of genomic distance in Section 1.5.2.

For natural graphs with an odd number of edges, it is impossible to complete them by adding pairs of gray edges. Nevertheless, as on the right of Figure 4.6, they may be grouped pairwise into *supernatural graphs* so that they may be completed with pairs of gray edges. Then, as with even natural graphs, 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.

Natural graphs with even numbers of edges are automatically supernatural graphs. Se is used to denote the number of this type of supernatural graphs. If there are natural graphs u and t both containing an odd number of edges, two such are joined to form a supernatural graph $s = u \cup t$. Priority is given to constructing supernatural graphs with four Os instead of two.

Let So^+ be the supernatural graph with four Os and B be the number of black edges in the breakpoints graph, then the genome halving distance is

$$d(A' \oplus A'', T) = \frac{B}{2} - Se - So^+ \tag{1.5.4}$$

The rules for drawing the gray edges can be found in [51].

1.5.4 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 we implement a heuristic using the principles of Bourque's MGR [26], but based on the constrained version of the Bergeron *et al.* [17] algorithm.

We assign a weight to a rearrangement operation based on the following rules: Let $RO(X)$ be a rearrangement operation for genome X. After we apply $RO(X)$ to get a new genome X', let $w_1 = d_r(X, Y) - d_r(X', Y)$ and $w_2 = d_r(X, Z) - d_r(X', Z)$, then $W_{RO(X)} = w_1 + w_2$.

The weight is used to choose the "right" rearrangement operation to apply at this stage, namely the one with the highest weight among all possible operations on X,Y or Z and that in addition best prepares for the next stage, according to a look-ahead routine. After the "right" operation (say without loss of generality, $RO(X)$) is chosen, we apply $RO(X)$ to get X', and reset X=X'. The strategy of [26] is to apply the "right" rearrangement operations to X, Y and Z recursively until three genomes are identical.

1.6 Overview of the chapters

In this thesis I propose several algorithms for the rearrangement phylogeny for sets of related genomes that include some that have undergone doubling. I investigate the properties of these algorithms and validate them by applying them to real data.

This thesis is based on eight papers that are published or in press. In six of these papers I was the lead researcher, in the eighth another co-author was primarily responsible although I made many substantive contributions, and in the fifth I contributed a section to a multi-authored review, which I nonetheless include because it provides a major justification for the thrust of my Ph.D. research.

In the six papers that are primarily my work, my supervisor Dr. David Sankoff was responsible for the English style, manuscript organization and some of the background writing, especially where he is listed as first author. The contributions of my other co-authors are detailed on the first page of each chapter.

Because each paper is originally a self-contained publication, there is some overlap and redundancy among them. This is especially true of the introductory and background sections of some of the papers, but these are not too lengthy and in each paper provide a quick review of what has been accomplished in the previous chapters. Fortunately, many of the papers share common terminology and notation, contributing to the coherence of the thesis as a whole. Finally, the bibliographies and lists of references have been consolidated to avoid duplication and to facilitate reading, and the chapters have been put into a common format, requiring some additional work on the figures and tables.

The ensuing chapters are organized as follows:

- **Chapter 2: Genome Halving with an Outgroup [238]**

  This was the first publication on what is now called "guided genome halving",

defining the problem and laying out a straightforward approach to its solution. For genomes that have incurred a genome doubling (tetraploidization) event in their evolutionary history, this event is reflected today in patterns of duplicated segments scattered throughout their chromosomes. These duplications may be used as data to "halve" the genome, i.e. to reconstruct the ancestral genome at the moment of tetraploidization, but the solution is often highly non-unique. To resolve this problem, I adapted the genome halving algorithm of El-Mabrouk and Sankoff [51] to take account of an external reference genome. The idea was basically to generate all halving solutions and to choose whichever one was closest to the reference genome. I applied this to reconstruct the tetraploid ancestor of maize, alternatively using rice or sorghum as the reference. This proved feasible because of the relatively small data set, involving only 34 segments occurring once in rice, once in sorghum and twice in maize, generating 1.5 million halving solutions to evaluate.

- **Chapter 3: Polyploids, genome halving and phylogeny [174]**

  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. I proposed a framework for solving this difficulty by integrating the rearrangement median and genome halving algorithms. I implemented a heuristic solution to the prototypical case of a tree with one tetraploid and two diploid genomes, and applied it to study the evolution of cereals and of yeast. Here, the exhaustive testing of all halving solutions still works for the cereal data with 34 homology sets, but would not be feasible for the yeast data containing 263 sets. Instead,

I devised a way of randomly sampling the space of halving solutions, and evaluated just a large sample, under the hypothesis that at least one of the samples would be close to the optimum.

- **Chapter 4: Descendants of whole genome duplication within gene order phylogeny [242]**

  A framework for phylogenetics on sets of diploid and polyploid genomes requires separating out various possible local configurations of doubled and unduplicated genomes in a given phylogeny. Each configuration requires a different strategy for integrating genomic distance, halving and rearrangement median algorithms. In this work I focused on the two cases where doubling precedes a speciation event and where it occurs independently in both lineages initiated by a speciation event. I applied these to a data set containing markers that are ancient duplicates in two yeast genomes, *Saccharomyces cerevisiae* and *Candida glabrata* to show that the tetraploidization occurred before speciation.

- **Chapter 5: Guided genome halving: hardness, heuristics and the history of the Hemiascomycetes [241]**

  I set out to improve on the computationally costly, exhaustive, method of Chapter 2 and the risky (and generally inaccurate) sampling method of Chapter 3. I adapted the genome halving algorithm of El-Mabrouk and Sankoff, by incorporating a look-ahead criterion for choosing among equally valid possibilities at each step, so that the algorithm rapidly and accurately constructs an ancestor close the outgroups, prior to a local optimization heuristic. The look-ahead is based on the new notion of "path groups". I applied this to reconstruct the predoubling ancestor of *Saccharomyces cerevisiae* and *Candida glabrata*, guided by the genomes of three other yeasts that diverged before the genome doubling

event. I analyzed the results in terms (1) of the minimum evolution criterion, (2) how close the genome halving result is to the final (local) minimum and (3) how close the final result is to an ancestor manually constructed by an expert with access to additional information. I also depicted graphically the set of reconstructed ancestors using classic multidimensional scaling to see what aspects of the two doubled and three unduplicated genomes influence the differences among the reconstructions.

- **Chapter 6: Polyploidy and angiosperm diversification [194]**

  Polyploidy has long been recognized a major force in angiosperm evolution. This chapter, reproducing a review co-authored under the leadership of D. Soltis, summarizes recent genomic investigations that not only indicate that polyploidy is ubiquitous among angiosperms, but also suggest several ancient genome doubling events. These include ancient whole genome duplication (WGD) events in basal angiosperm lineages, as well as a proposed paleohexaploid event that may have occurred close to the eudicot divergence. The question is no longer what proportion of angiosperms are polyploid, but how many episodes of polyploidy characterize any given lineage. A major challenge that biologists now face is to reconstruct the ancestral genomes of lineages prior to genome duplication. New algorithms such as the ones I describe in Chapters 2-5, and illustrated here, provide promise that ancestral genomes can be reconstructed over long periods of evolutionary time (e.g., it may be possible to reconstruct the ancestral eudicot or even the ancestral angiosperm genomes). Comparisons of diversification rates in this paper also suggest that genome doubling has lead to a dramatic increase in species richness in a number of angiosperm lineages, including Poaceae, Solanaceae, Fabaceae, and Brassicaceae.

- **Chapter 7: Gene loss under neighbourhood selection following whole genome duplication and the reconstruction of the ancestral Populus diploid [244]**

  Branching out from purely structural considerations, I developed criteria to detect neighbourhood selection effects on gene loss following whole genome duplication, and applied them to the (*Populus trichocarpa*) genome. I also continued the improvement of the "path groups" approach (Chapter 5) to guided genome halving algorithms so that more than two thousand gene sets, each containing two paralogs in the descendant T of the doubling event and their single ortholog from an undoubled reference genome R, can be analyzed to reconstruct the ancestor A of T at the time of doubling. At the same time, large numbers (up to 4000) of defective gene sets, either missing one paralog from T or missing their ortholog in R, could be incorporated into the analysis in a consistent way. I applied this approach to the poplar and grapevine (*Vitis vinifera*) genomes, as T and R respectively.

- **Chapter 8: Towards improved reconstruction of ancestral gene order in angiosperm phylogeny [175]**

  Because the path groups approach to guided genome halving used in Chapters 5 and 7 contains an inherent bias favouring one of the two tetraploid gene orders over the reference diploid order, I devised an unbiased version of the algorithm. This adds to the numerous questions about method and data that have cropped up in this approach to guided genome halving. Thus I investigated a battery of techniques and data choices, including the new unbiased algorithm, reference to two related genomes instead of only one to guide the reconstruction, use of draft sequences in contig form only, incorporation of incomplete sets of

homology correspondences among the genomes and addition of large numbers of "singleton" correspondences. I made use of genomic distance, breakpoint reuse rate, dispersion of sets of alternate solutions and other means to evaluate these techniques and choices, while reconstructing the pre-WGD ancestor of *Populus trichocarpa* as well as an early rosid ancestor. I used both *Vitis vinifera* and *Carica papaya* as reference genomes.

- **Chapter 9: Multichromosomal median and halving problems under different genomic distances [208]**

  Genome median and genome halving are related combinatorial optimization problems that aim at reconstructing ancestral genomes as well as the evolutionary events leading from the ancestor to extant species. Exploring computational complexity issues is an important adjunct to devising efficient algorithms.

  The complexity of the median problem for unichromosomal genomes (permutations) had previously been settled for both the breakpoint distance and the reversal distance. Although the multichromosomal case was often assumed to be a simple generalization of the unichromosomal case, technically it is a "relaxation", so that complexity in this context does not follow from existing results, and is open for all genomic distances. In this paper, under the research leadership of E. Tannier, inspired by my paper contained here as Chapter 5, we settled the complexity of several genome median and halving problems, including a surprising polynomial result for the breakpoint median and guided halving problems in genomes with circular and linear chromosomes, showing that the multichromosomal problem is actually easier than the unichromosomal problem. Still other variants of these problems are NP-complete, including the so-called DCJ double distance problem. Our theoretical study cleared up a wide swathe

of the algorithmical study of genome rearrangements with multiple multichromosomal genomes and clarified the relations among a number of variants of multichromosomal genome rearrangement problems.

Chapter 2

Genome halving with an outgroup

Chunfang Zheng, Qian Zhu and David Sankoff. 2006. Evolutionary Bioinformatics 2:319–326.

I was responsible for designing the algorithm and getting all of the results in this work. I also collaborated fully in writing and preparing the manuscript, especially the technical write-ups and graphics. Qian Zhu identified the syntenic blocks data of maize, rice and sorghum based on the information provided by Gramene [82].

Abstract

Some genomes are known to have incurred a genome doubling (tetraploidization) event in their evolutionary history, and this is reflected today in patterns of duplicated segments scattered throughout their chromosomes. These duplications may be used as data to "halve" the genome, i.e., to reconstruct the ancestral genome at the moment of tetraploidization, but the solution is often highly non-unique. To resolve this problem, we adapt the genome halving algorithm of El-Mabrouk and Sankoff to take account of an external reference genome. We apply this to reconstruct the tetraploid ancestor of maize, using either rice or sorghum as the reference.

2.1 Introduction

Many genomes have been shown to result from an ancestral doubling, or tetraploidization, event, followed by a period of diploidization, i.e., the loss of compartmentalization between the two original copies of the genome, as well as genome rearrangement through intra- and interchromosomal movement of genetic material. The genome halving problem is to reconstruct the ancestral genome on the basis of a decomposition of the present-day genome into a set of apparently duplicated blocks of genes or DNA sequence dispersed among the chromosomes. A quantitative approach to this problem was first discussed by Seoighe and Wolfe [180] in the context of the genome doubling of the ancestor of the yeast *Saccharomyces cerevisiae*. At the same time, motivated by studies of genome duplication in early vertebrates [143], El-Mabrouk and colleagues [47, 48, 50] published a series of papers on the combinatorial optimization approach to the problem, culminating in a general solution [51]. Further refinements have been published by Alekseyev and Pevzner [6].

Seoighe and Wolfe [180] noted the extreme non-uniqueness associated with the solution to the genome halving problem and suggested that this difficulty could be attenuated through the use of a reference genome, or outgroup. The suggestion to use a reference genome was taken up to study the post-tetraploidization evolution of *S. cerevisiae*, both in reference to the genome of *Ashbya gossypii* [44] and to that of *Kluyveromyces waltii* [86], though without recourse to genome rearrangement or genome halving algorithms. Similar research compared mammalian genomes with the tetraploid ancestor of the pufferfish *Tetraodon nigroviridis* [80]. In the present paper, we formalize this strategy by developing a general algorithm to reconstruct an ancestral tetraploid genome with reference to an outgroup genome. We apply it to infer the ancestor of the maize (*Zea mays*) genome, with the rice (*Oryza sativa*)

and sorghum (*Sorghum bicolor*) genomes as outgroups. 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 [82], and not the rest of the genomes.

Our strategy is to generate all the solutions to the genome halving problem for the maize genome, and to focus on the subset of these that have a minimum rearrangement distance with the rice (or sorghum) genome. We formulate a search heuristic to transcend the set of optimal halving solutions to find the most realistic ancestral genome that minimizes the sum of the distance between the ancestral tetraploid and present-day maize and the distance between rice (or sorghum) and the diploid form of the ancestor.

2.2 The Data

It is generally agreed that the maize genome underwent a genome doubling event some 11-16 million years ago [60]. While some duplicated regions clearly attest to this event, there is no consensus on the exact inventory of such regions. Moore *et al.* and Wilson *et al.* [137, 225] presented two largely consistent views of syntenic blocks across the cereals based on the mapping evidence at the time. These included 14 and 19 duplicated blocks in the maize genome. Gaut [61] gave a more comprehensive account of the pattern of 23 duplicated regions, based on maize genomic sequence data in 2001. He did not completely establish the relative position of all the syntenies on the chromosomes in this work.

Even now that the rice genome has been sequenced, and the maize genome project is well-advanced, it is no trivial matter to identify the duplicate blocks resulting from the tetraploidization event. The maize genome has many other duplicated segments

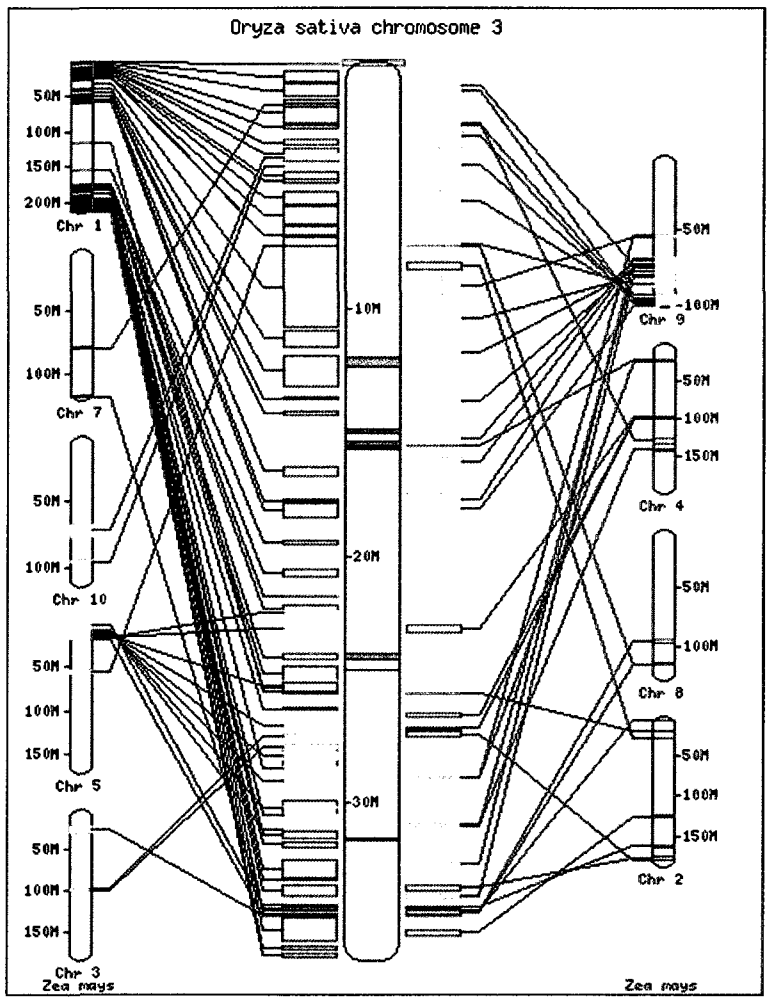

Figure 2.1: Syntenies between rice chromosome 1 and maize chromosomes, as produced by Gramene [82]

dating from periods both after and before the tetraploidization and even before the divergence from the other cereals. This is complicated by post-tetraploidization genome rearrangement events, deletions and insertions of genetic material, transpositions of genes or larger segments from one site on the genome to another, and loss of homology between the parts of the duplicated regions.

The databank which has the most information on the syntenies among the cereal genomes is Gramene [82]. The current version at time of writing is release 21.

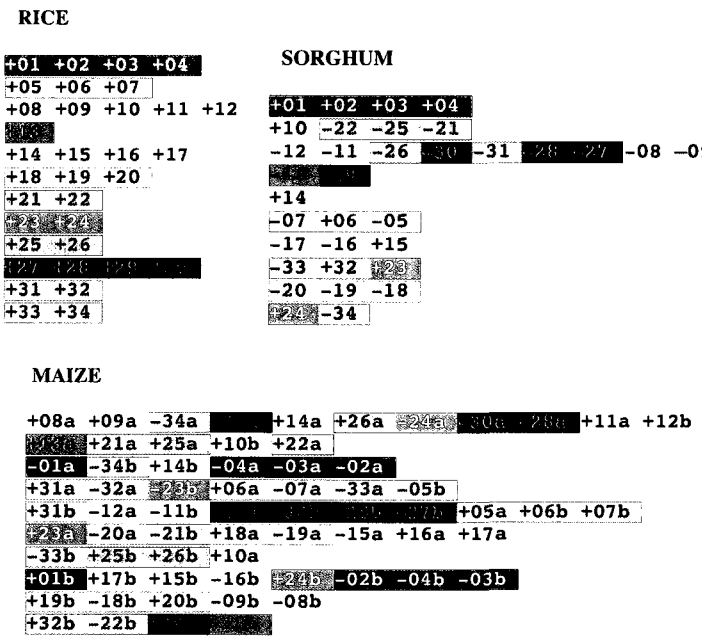

Figure 2.2: Order of syntenic blocks in rice, sorghum and, in two copies each, maize.

From this we can obtain a conservative (i.e., confined to high homology regions only) estimates of duplicate blocks by comparison with the rice genome. For example, in Figure 2.1, we can visually identify large duplicated regions in chromosomes 1 and 9, chromosomes 1 and 5, and possibly a number of smaller ones, all by virtue of their common homology with regions of rice chromosome 3.

Unfortunately, there is as yet no comparison of syntenic blocks between sorghum and the other genomes on Gramene. However, there are extensive mapping data of various kinds of markers. We bolstered our preceding data collection by searching sets of duplicate markers in maize that had single copies in sorghum and rice, comparing mainly the Patterson 2003 genetic map for sorghum, the IBM2 Neighbours 2004 and Cornell Wilson 1999 genetic maps for maize and the Annotated Nipponbare Sequence 2006 sequence map for rice. All the markers satisfying these criteria fell into the rice-maize syntenies established previously. Based on these criteria, i.e., markers identified

as homologous in Gramene, with a single copy in each of rice and sorghum and two copies in maize, plus the requirement that the maize and rice copies fall into the appropriate, previous identified, rice-maize syntenic blocks, we could now identify 34 syntenic blocks as basic data for our reconstruction. These data are depicted in Figure 2.2, but should be considered to constitute a working hypothesis; definitive data must await the finishing of the maize genome, the sequencing of the sorghum genome, and the further application of global alignment and synteny block construction methods.

2.3 The Genome Halving Algorithm

Distance based on genomic structure $d(X, Y)$ is calculated by rapid, albeit complicated, rearrangement algorithms for finding the minimum number of operations necessary to convert one genome X into another Y. The genomes are represented by signed permutations on $1, \cdots, n$ and the biologically-motivated operations generally include inversions (implying as well change of sign, i.e., change of strand) of chromosomal segments, reciprocal translocations (of telomere-containing segments of two chromosomes) and chromosome fission or fusion. They may also include transpositions (including "jumping genes") of segments from one site to another on a chromosome or interchanges of segments on a chromosome, both of which count as two steps compared to one for the previously mentioned operations.

Rearrangement algorithms (e.g., [210]) make use of the bi-coloured "breakpoint graph" or similar structure, where each end of an oriented syntenic block, gene or marker on genome X is joined by a red edge to the adjoining end of the adjacent syntenic block, gene or marker, and these same ends, represented by the $2n$ vertices in the graph are joined by blue edges determined by the adjacencies in genome Y. The breakpoint graphs necessarily consist of disjoint alternating cycles and/or paths,

and it can be shown that $d(X,Y) = n - c$, where c is the number of cycles (in the case X and Y consist of single circular chromosomes), or $d(X,Y) = n + \chi - c - \Pi$, where χ is the maximum number of linear chromosomes in X and Y, and Π counts the number of certain kinds of paths in the graph. The actual operations, $d(X,Y)$ in number, may be reconstructed by splitting large cycles in the breakpoint graph into two cycles each, until there are $d(X,Y)$ cycles each made up of two vertices, one red edge and one black edge. Every time a cycle is split, this corresponds to one rearrangement operation.

In the rearrangement algorithms, construction of the breakpoint graph is an easy preliminary step. The genome halving algorithms [6, 51] also make use of the breakpoint graph, but the problem here is building the breakpoint graph where one of the genomes (the tetraploid) is unknown. This is done by segregating the vertices of the graph in a natural way into subsets, such that the vertices of all cycles must fall within a single subset, and then constructing these cycles in an optimal way within each subset so that the red edges correspond to the structure of the known genome and the black edges define the adjacencies of a tetraploid.

2.4 A Heuristic for Minimizing $d(U,A) + d(A \oplus A, T)$

Let T be a genome consisting of χ chromosomes and $2n$ genes, syntenic blocks or other markers, $g_{1,1} \cdots, g_{1,n}; g_{2,1}, \cdots, g_{2,n}$, dispersed in any order on the chromosomes. For each i, we call $g_{1,i}$ and $g_{2,i}$ "duplicates", but there is no particular property distinguishing elements of the set of $g_{1,i}$ from the set of $g_{2,i}$. A potential "ancestral tetraploid" of T is written $A \oplus A$, and consists of 2ψ chromosomes, where some half (ψ) of the chromosomes contains exactly one of each of $g_{1,i}$ or $g_{2,i}$ for each $i = 1, \cdots, n$. The remaining ψ chromosomes are each identical to one in the first

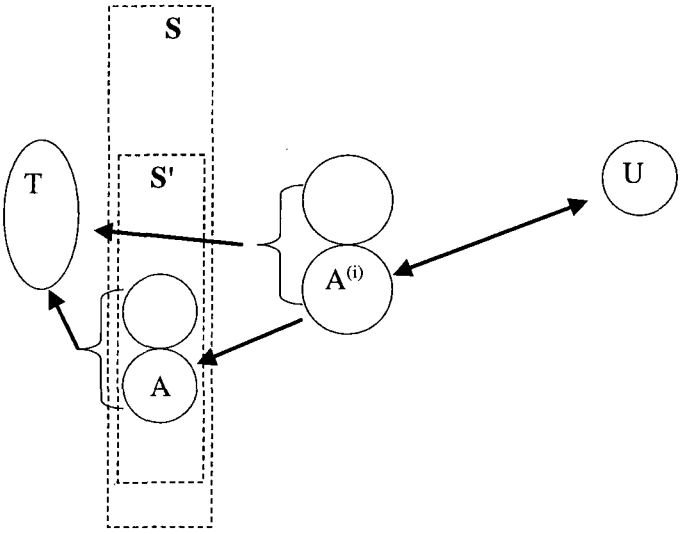

Figure 2.3: Procedure for finding ancestral tetraploid. T=genome made up of duplicated markers, U = reference genome. $\mathbf{S}$ = set of solutions to the genome halving problem. $\mathbf{S}'$ = subset closest to U, $A^{(i)}$ = genome on trajectory from $A \in \mathbf{S}'$ to U.

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 where $g_{2,i}$ appears in the first half, $g_{1,i}$ appears in the second. 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 ψ chromosomes, and the n genes, syntenic blocks or markers they contain, $g_1, \cdots, g_n$ 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.

Any genome U is a reference genome for T if it contains the n genes, syntenic blocks or markers $g_1, \cdots, g_n$.

Let U be a reference genome for T. The central problem in this paper is to find a potential ancestral diploid genome A such that $d(U, A) + d(A \oplus A, T)$ is minimized.

Let $\mathbf{S}$ be the set of solutions of the genome halving algorithm for T. As an initial step to our heuristic, schematized in Figure 2.3, we confine our search to $\mathbf{S}$.

For each solution $A \in \mathbf{S}$, we calculate the rearrangement distance $d(U,A)$ between the reference genome U and A. This is feasible even for large $\mathbf{S}$ because of the rapidity of the rearrangement calculation. We then define

$$\mathbf{S}' = \{A \in \mathbf{S} | d(U,A) = \min_{X \in \mathbf{S}} d(U,X)\}. \tag{2.4.1}$$

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

$$d(U, A^{(i)}) + d(A^{(i)} \oplus A^{(i)}, T) < d(U,A) + d(A \oplus A, T). \tag{2.4.2}$$

(Note that it is not necessary to try $A^{(1)}$ though it is closer by one step to U than A is, because $A^{(1)} \oplus A^{(1)}$ is also farther from T by at least one step, since it is not in $\mathbf{S}$.) Our final solution set $\mathbf{S}''$ is the set of $A^{(i)}$, over all genomes $A \in \mathbf{S}'$, and all trajectories between A and U, that satisfy inequality (2.4.2) and that minimize the left hand side of (2.4.2).

If $\mathbf{S}''$ is empty, then $\mathbf{S}'$ is the final set of minimizing genomes.

2.4.1 Complexity

Since both genome halving and genome rearrangement are essentially linear in n, the execution time of our search is $O(n|\mathbf{S}| + \phi n^2 |\mathbf{S}'|)$, the second term measuring the number of steps between genomes in $\mathbf{S}'$ and U and the time to calculate the distance to U at each step, and the number ϕ of different paths sampled per element

in $\mathbf{S}'$. In our example, biological reality motivates constraining the search so that all chromosome fissions are carried out first, as far as compatible with the optimality of the path. This is because the loss of chromosomes is likely to occur around the time of diploidization, so the path back from A towards the ancestor should attempt to restore the number of chromosomes to what it is in sorghum or rice as soon as possible, i.e., for some $A^{(i)}$, where i is as small as possible.

2.5 Results

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 5,000,000 different execution paths for the algorithm. Not all of these lead to the same results, but distinct solutions in $\mathbf{S}$ surely number in the hundreds of thousands, if not millions; a sample of 15,000 paths resulted in over 13,000 different solutions. 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 reference genomes to bear, we first note that over all $X \in \mathbf{S}$, the distance $d(X, So)$ ranges from 19 (for the solutions in $\mathbf{S}'_{\mathbf{So}}$) to 28, while $d(X, R)$ ranges from 19 (for the solutions in $\mathbf{S}'_{\mathbf{R}}$) to 27. The sets $\mathbf{S}'_{\mathbf{So}}$ and $\mathbf{S}'_{\mathbf{R}}$, however, contain only 8 and 24 solutions, respectively. Thus there is a massive reduction of non-uniqueness induced by appealing to a reference. Then, in venturing outside of $\mathbf{S}$ on the paths from pre-tetraploid versions of elements of $\mathbf{S}'$ towards the reference, either rice or sorghum, we find even fewer genomes X with a minimum sum of distance to the reference (X as a diploid) plus distance to maize ($X \oplus X$ as a tetraploid). For

example, the genome $A_{SO}^{(3)}$ in Figure 2.4 and depicted in Figure 2.5 satisfies

$$\begin{aligned} d(So, A_{SO}^{(3)}) + d(A_{SO}^{(3)} \oplus A_{SO}^{(3)}, M) &= 16 + 29 \\ &< d(So, A) + d(A \oplus A, M) \\ &= 19 + 27, \end{aligned} \tag{2.5.1}$$

for all $A \in \mathbf{S}'$. There are only two other solutions with value 45 for the objective function, one a step closer (an $A^{(4)}$) and one a step further (an $A^{(2)}$), from the sorghum genome. In the case of a rice reference, there is actually a unique solution, with $d(R, X) + d(X \oplus X, M) = 44$.

Thus we have almost completely eliminated the non-uniqueness of the solutions to the genome-halving problem, though of course the number of solutions found will still depend on the data set. It is also possible that a better solution is to be found off the paths we have explored, although this is unlikely for the relatively small example represented by these cereal genomes.

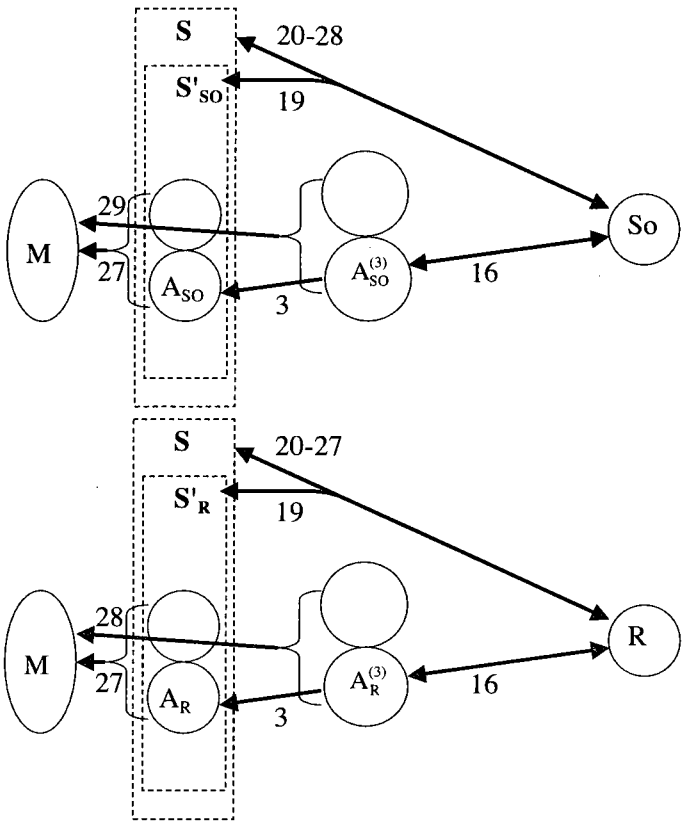

Figure 2.4: Results of search for ancestral tetraploid. M, R, SO = maize, rice,sorghum genomes. $\mathbf{S}$ = set of solutions to the genome halving problem. $\mathbf{S'_R}, \mathbf{S'_{SO}}$ = subsets closest to R, SO, $A_R^{(i)}$ = genome on trajectory from $A_R \in \mathbf{S'_R}$ to R. $A_{SO}^{(i)}$ = genome on trajectory from $A_{SO} \in \mathbf{S'_{SO}}$ to SO.

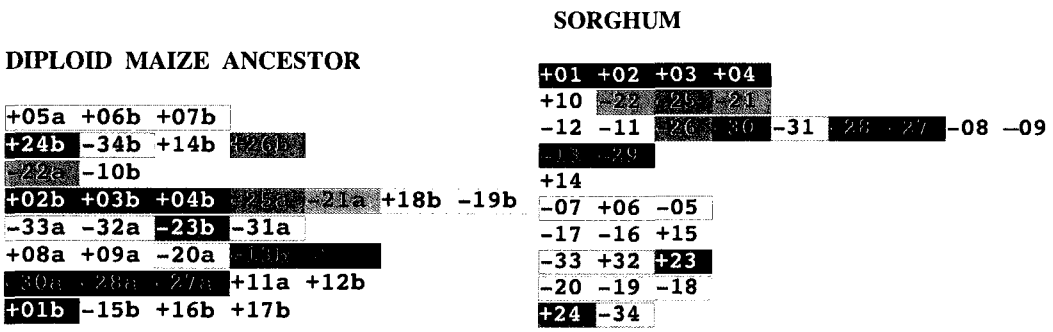

Figure 2.5: Order of syntenic blocks in the reconstructed diploid maize ancestor, compared to sorghum, with the same rice chromosomal colour coding as in Figure 2.2.

2.6 Conclusions

We have been working with a small data set, and the differences between the optimal solution and suboptimal solutions are small, as in inequality (2.5.1). As more data become available on maize and especially sorghum, our reconstructions should be better and the role of the reference genome in zeroing in on a unique solution for genome halving will be clarified. This should also allow for statistical validation.

Our analysis used sorghum and rice as reference genomes in two separate analyses. And it is gratifying that using sorghum alone as reference produced an ancestral maize genome closer, not only to sorghum, but also to rice, than any candidate ancestor based on genome halving with no reference. Nevertheless, it would be interesting to formally combine gene order information from both rice and sorghum simultaneously in reconstructing the maize ancestor. Along the lines of our current analysis, first finding $\mathbf{S}$, then $\mathbf{S}'$, and finally an optimal $A^{(i)}$, we could define $\mathbf{S}'$ as the subset of $\mathbf{S}$ whose elements A each induce a minimal solution of the median problem [169,181], i.e., for which there is a genome X, such that $d(A,X)+d(U_1,X)+d(U_2,X)$ is minimal compared to all $A \in S$. Then the search for an optimal $A^{(i)}$ could proceed on the

paths from all $A \in S'$ to X.

A more difficult theoretical problem would be to replace our sequential procedure by a single algorithm searching for the A which minimizes $d(U, A) + d(A \oplus A, T)$. It is not clear whether this is a hard problem, given that genome halving and genome rearrangement are both solvable in close to linear time. But there is no obvious way of modifying the halving algorithm so that it could take account of a reference genome while retaining optimality. Some of the searches we have performed here might be incorporated directly into the halving algorithm to transform it into a heuristic method, and this might work even for $d(U_1, A) + d(U_2, A) + d(A \oplus A, T)$.

Chapter 3

Polyploids, genome halving and phylogeny

David Sankoff, Chunfang Zheng and Qian Zhu. 2007. Bioinformatics 23:i433–i439.

I was responsible for designing the algorithm and getting all of the results in this work. I also collaborated fully in writing and preparing the manuscript, especially technical write-ups and graphics. Qian Zhu extracted the yeast data from the literature [44, 86].

Abstract

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.

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.

3.1 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 paper 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 3.2, 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 neighbouring genomes in a binary branching tree. In

[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.

Section 3.3 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 3.4 and 3.5, we apply our method to a small data-set on maize and a large data-set on yeast, respectively.

3.2 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 3.1. 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 phy-

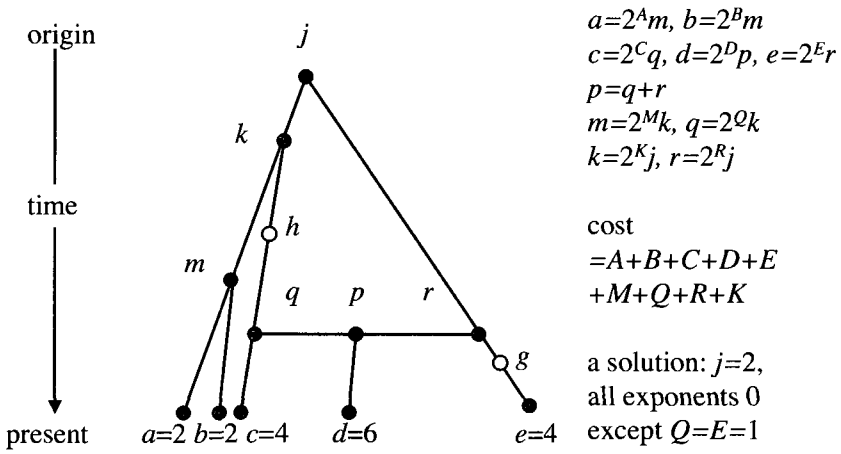

Figure 3.1: Example of ploidy inference problem. Genomes labeled by ploidies, observed only for leaves of phylogeny. Tetraploidy events inferred at g and h, or alternatively on the branches jr and qc.

logenetic 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 paper, 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 3.1, 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.

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 3.3.1 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 labelled a and b in Figure 3.1, 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 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 labelled m in Figure 3.1? 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 3.3.1.

3.3 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.

3.3.1 Existing and missing resources

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 [17], which we constrain to allow only the operations we have listed.

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 [51].

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 [26] MGR (multiple genome rearrangement) algorithm, but based on the constrained version of the [17] distance algorithm.

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 3.3.2 below, we would have to generalize the genome halving problem in several direction:

- 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.

3.3.2 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, $g_{1,1} \cdots, g_{1,m}; g_{2,1}, \cdots, g_{2,m}$,

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 $i = 1, \cdots, m$. 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 syntenic blocks they contain, $g_1, \cdots, g_m$, 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 $g_1, \cdots, g_m$, as in Figure 3.2a. Our central problem is to find a diploid genome A and a median genome M of A, R_1 and R_2 that minimize

$$D(T, R_1, R_2) = d(R_1, M) + d(R_2, M) + d(A, M) + d(A \oplus A, T). \tag{3.3.1}$$

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 3.2b, 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 3.2c. This is the bottleneck in our procedure, since $\mathbf{S}$ may be very large, and an accurate calculation of the median is costly

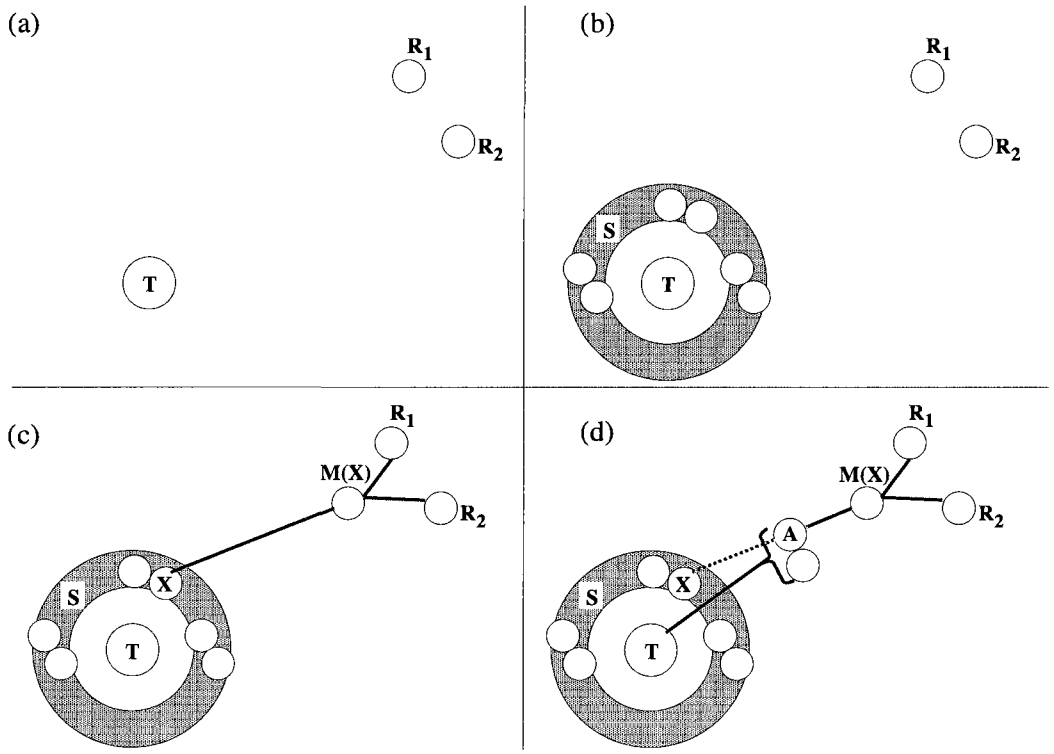

Figure 3.2: Strategy for phylogenetically-constrained genome halving. (a) Descendant T of ancestral tetraploid, with two related diploids R_1 and R_2. (b) Set $\mathbf{S}$ of solutions of genome halving of T, showing pairs of fused identical diploids. (c) Solution $X \in \mathbf{S}$ that also induces 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 each element of $\mathbf{S}$. When the number of markers m is small, say a few dozen, as to be illustrated in Section 3.4 below, it is possible to do a complete evaluation of $\mathbf{S}$. When m is in the hundreds, as to be illustrated in Section 3.5 below, we resort to a random sample of the genomes in $\mathbf{S}$.

We then define

$$\begin{aligned} \mathbf{S}' &= \{X \in \mathbf{S} | d(R_1, M(X)) + d(R_2, M(X)) + d(X, M(X)) \\ &\qquad \text{is a minimum}\}. \end{aligned} \tag{3.3.2}$$

By definition, there is no minimizing genome in $\mathbf{S} \setminus \mathbf{S}'$. 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 3.2d. 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

$$\begin{aligned} d(X^{(i)}, M(X)) + & d(X^{(i)} \oplus X^{(i)}, T) \\ & < d(X, M(X)) + d(X \oplus X, T). \end{aligned} \tag{3.3.3}$$

For relatively small examples, e.g., for the data in Section 3.4, we can also iterate on the median step, and look for

$$\begin{aligned} d(X^{(i)}, M(X^{(i)})) + & d(X^{(i)} \oplus X^{(i)}, T) \\ & < d(X, M(X)) + d(X \oplus X, T). \end{aligned} \tag{3.3.4}$$

Any genome $X^{(i)}$ that minimizes the left hand side of inequality (3.3.3) or, better,

inequality (3.3.4), 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.

3.4 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 [60]. 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 [82], and not the remainder of each of the genomes.

In a previous study [238] 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 $\mathbf{S}$. 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.

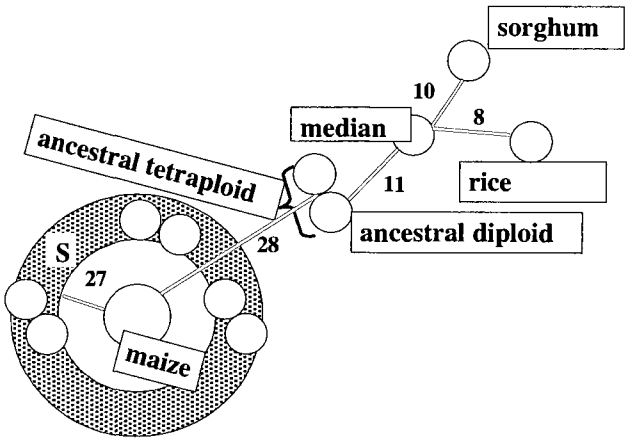

Figure 3.3: Solution for the maize data.

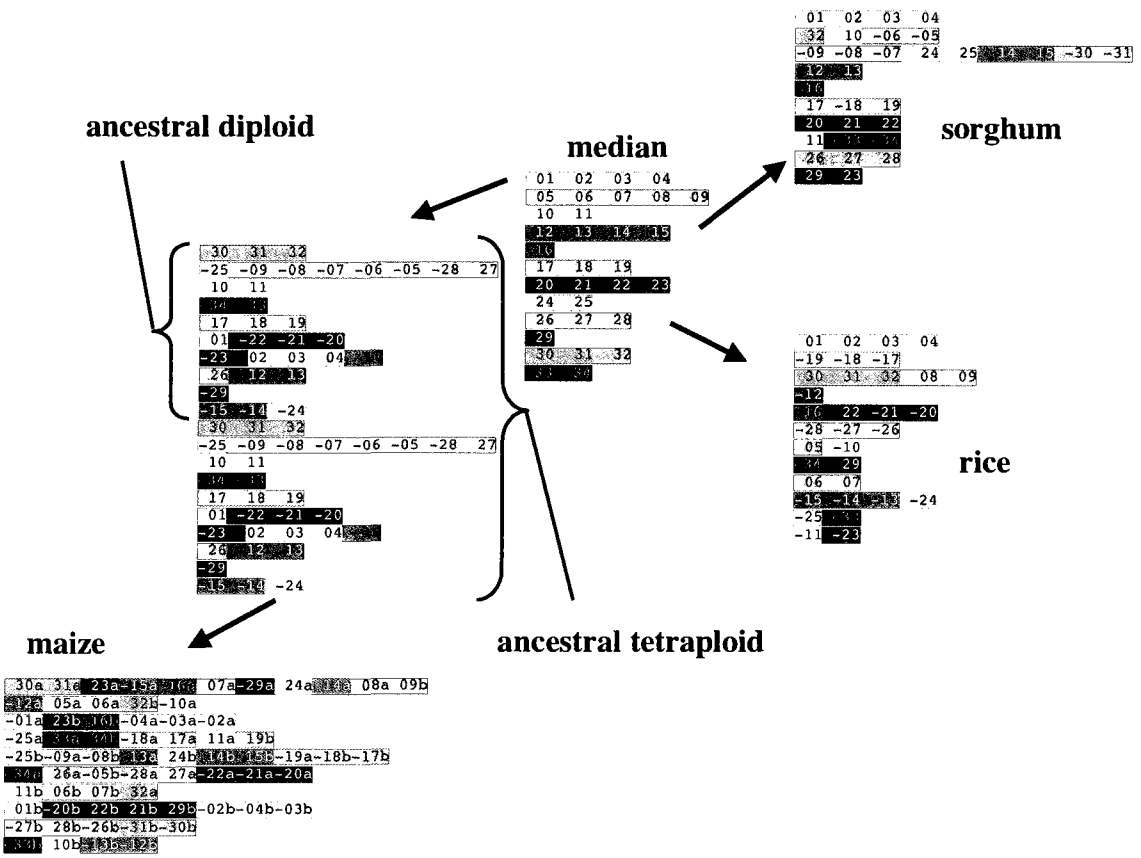

Figure 3.4: Given and inferred cereal karyotypes and synteny blocks, colour-keyed to the median genome.

When we bring the diploid genomes to bear using Equation (3.3.2), however, testing all 1,500,000 elements of $\mathbf{S}$, the set $\mathbf{S}'$ contains only 9 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 $\mathbf{S}'$ towards the median using the criterion in inequality (3.3.4) led directly to the solution in Figure 3.3, which is not improved by local searching. Other trajectories from $\mathbf{S}'$ 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 3.3, the given and inferred genomes, with synteny blocks evident, are depicted in Figure 3.4.

3.5 Tetraploidization of yeast

Wolfe and Shields [227] 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* [44] and of *Kluyveromyces waltii* [86], though without recourse to genome rearrangement or genome halving algorithms.

Each of these studies located 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.

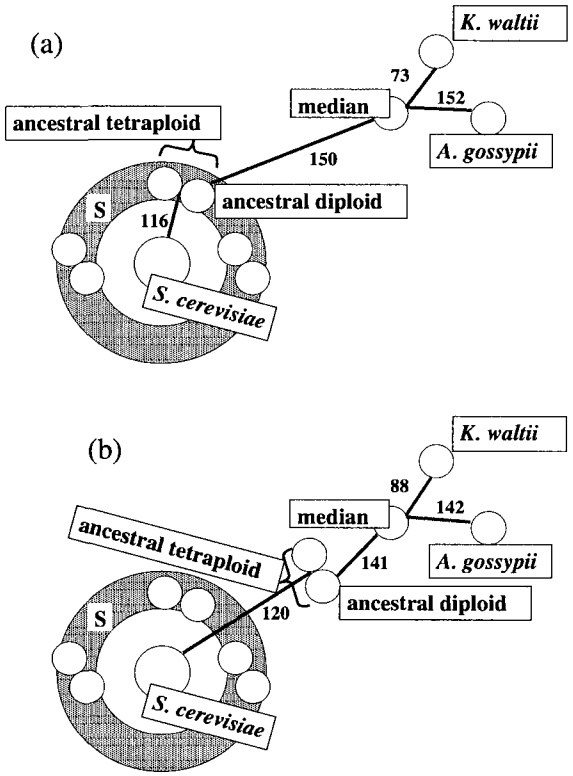

Figure 3.5: Two solutions for the yeast data. (a) Solution $\in \mathbf{S}$. (b) Solution $\notin \mathbf{S}$ and detailed in Figure 3.7.

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 3.5. 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 $\mathbf{S}$ with $m = 263$, and found an $\mathbf{S}'$ with only one element.

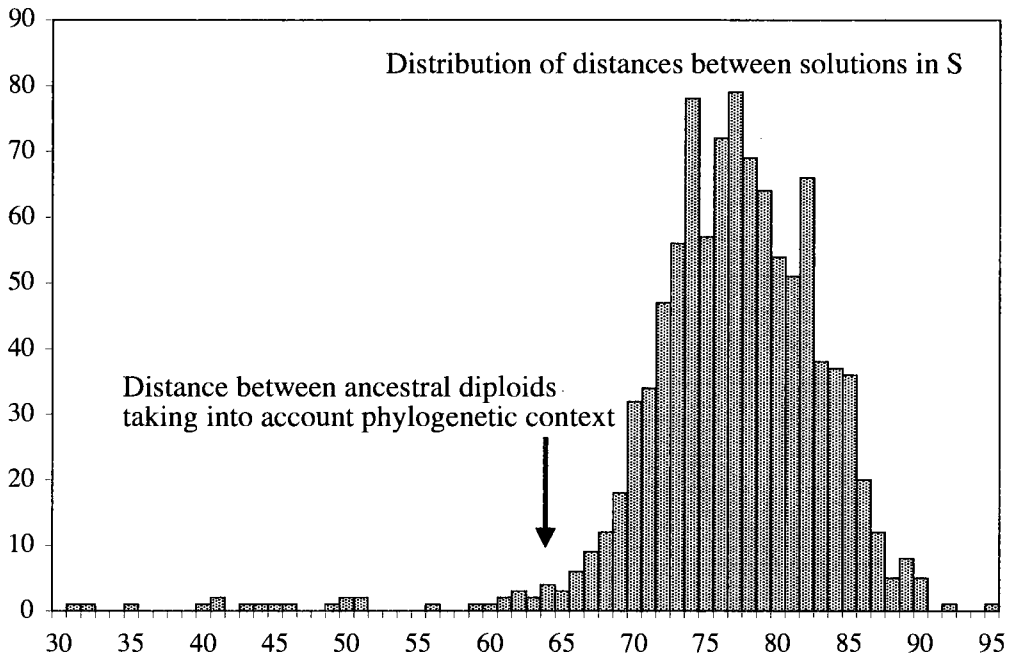

Figure 3.6: Distribution of distances between genomes in $\mathbf{S}$

To compensate for the sketchy coverage of $\mathbf{S}$, we also examined several solutions of the genome halving algorithm where D was slightly sub-optimal. Furthermore, we used the criterion in inequality (3.3.3) instead of the computationally more costly inequality (3.3.4) 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 $M(B)$.

How different are these two solutions, summarized in Figure 3.5? If we calculate the rearrangement distance between them and compare it with randomly chosen pairs of genomes in $\mathbf{S}$ as in Figure 3.6, 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.

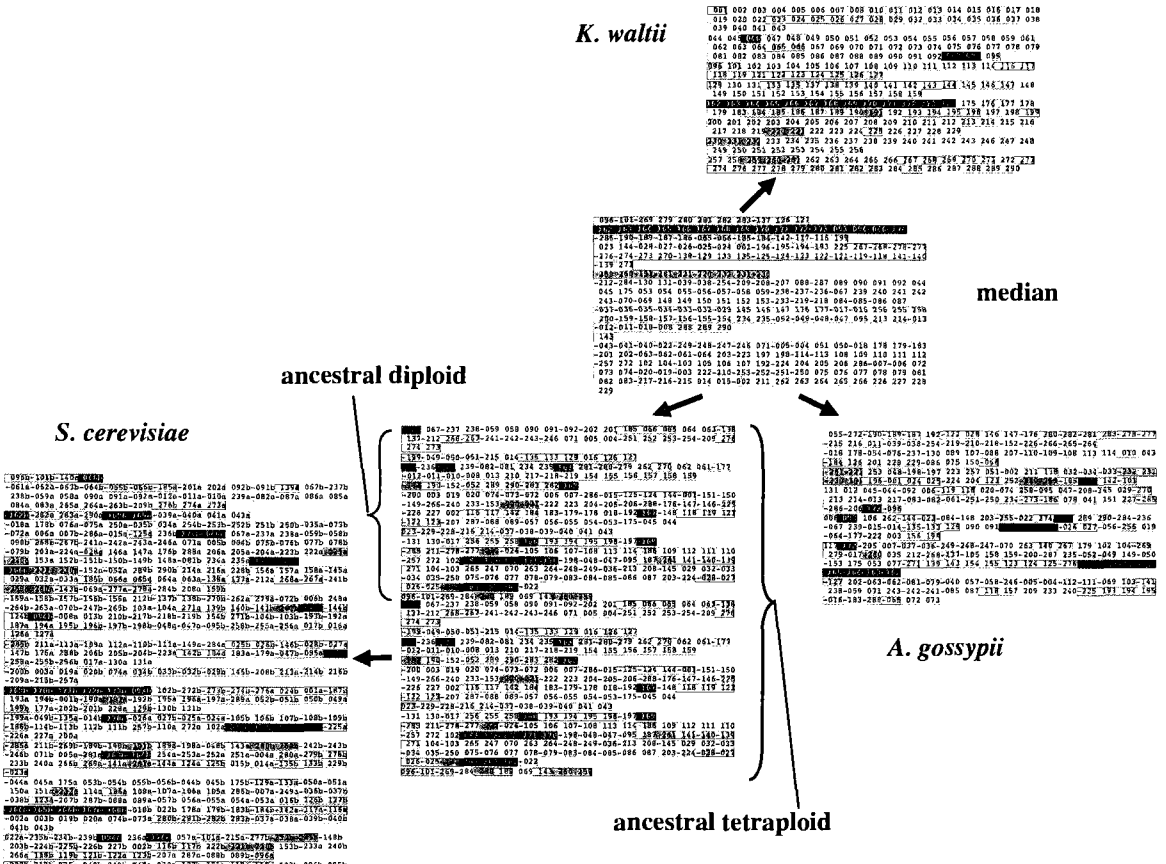

Figure 3.7: Given and inferred yeast karyotypes and synteny blocks, colour-keyed to the median genome. Long chromosomes are wrapped and chromosomes are separated by thin white space.

3.6 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 3.3, 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 $\mathbf{S}$, but would not mitigate the complexity of the median step.

Acknowledgement

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

Chapter 4

Descendants of whole genome duplication within gene order phylogeny

Chunfang Zheng, Qian Zhu and David Sankoff. 2008. Journal of Computational Biology 15:1–18. [Conference version: Chunfang Zheng, Qian Zhu and David Sankoff. 2007. Parts of the problem of polyploids in rearrangement phylogeny. in Tesler G, Durand D, eds. RECOMB 2007 Workshop on Comparative Genomics. LNCS 4751:162–176, Springer.]

I was responsible for designing the algorithm and getting all of the results in this work. I also collaborated fully in writing and preparing the manuscript, especially technical write-ups and graphics. Qian Zhu extracted the yeast data from YGOB [31].

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.

4.1 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. Evidence for the effects of genome duplication has shown up across the eukaryote spectrum. Aside from the well-known controversy about doubling in vertebrates more than two hundred million years ago, e.g., [73, 134], more recent genome duplications are known to have occurred in some vertebrate lines, such as the the zebrafish [157], the salmoniform fish [144], frogs [229] and even mammals, as evidenced by the rats *Tympanoctomys barrerae* [55] and *Pipanacoctomys aureus* [56]. Genome duplication is widespread in insects and particularly prevalent in plants (see [41] for a survey of the flowering plants).

For some generations after a true doubling event (called autotetraploidization in classical genetics), the meiotic process, characterized by the normal bivalent pairings of one maternal and one paternal chromosome, may be disrupted by abnormal quadrivalent or trivalent alignments involving homeologous chromosomes, singleton chromosomes and other aberrant structures, resulting in reduced fertility. Doubling is followed by a period of "re-diploidization", where distinct pairings again emerge, though in approximately twice the original number, a process mediated by sequence divergence and by chromosomal rearrangement, through intra- and interchromosomal movement of genetic material. These rearrangement processes continue after diplo-

somy is attained, so that eventually the chromosomal neighbourhood of a marker need bear no resemblance to that of its duplicate.

Doubling may also occur as a fusion of two distinct but related genomes (allotetraploidy) instead of autotetraploidy. With allotetraploidy, the period of unpredictable meiotic patterns may be absent or attenuated, to the extent that meiotic pairing remains true to the parental types, and for this reason it is thought that this process has a better chance of leading to viable offspring. Intra- and interchromosomal movement of genetic material will also eventually result in the karyotypic mixture of the two founding genomes after this type of doubling.

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.

For descendants of autotetraploids, 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. In the case of allotetraploids, if there are data on unduplicated species more closely related to one of the contributing genomes than the two contributors are to each other, this may allow systematically partitioning the pairs into two sets accordingly. The mathematics of this case are different [49], and will not be discussed here. When there are no intervening unduplicated genomes, it may not even be possible to detect that this was an allotetraploid rather than an autotetraploid and there will be little distinction to be made on the practical level for the analysis of the two cases.

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, or unresolved two-to-two relationships between two doubling descendants,

whereas these algorithms require a one-to-one correspondence.

We have undertaken a program [174, 238] 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 4.2, we outline a model for generating an arbitrary pattern of doubling 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 4.3, we identify the three basic algorithms underlying the study of genome rearrangements and gene order phylogeny, 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 ancestor 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 4.4, we review the methodology for the small phylogeny problem using gene order data, based on the iterative application of the median algorithm successively to all the ancestral vertices of a phylogenetic tree. When doubling descendants

are considered together with unduplicated genomes, there are four kinds of median problem; these are reviewed in Section 4.5.1. In Section 4.5.3, we discuss our recent papers [174, 238] on incorporating the three basic algorithms: distance, halving and median, 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 4.6 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. To systematically assess the biases in these approaches, in Section 4.7, we carry out a series of simulations. In Section 4.8, we apply these two methods to a large data set on yeast.

4.1.1 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 many 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. Whether these are haploid, diploid or some other ploidy is not relevant in that in any organism that reproduces

by meiosis, the order of the markers on any of the aligned 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.

4.2 Inference of doubling events

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 mapping information is much easier to obtain experimentally and to come by in the literature than complete sequence information, though the nature and timing of 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 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 at each step, and the parsimony criterion that the number of doubling steps is to be minimized, the

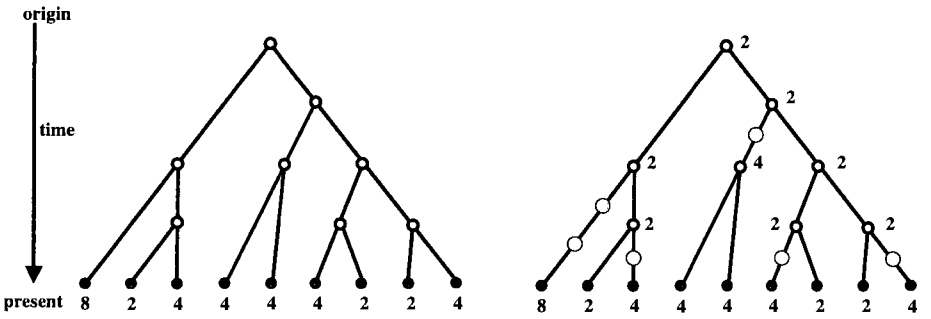

Figure 4.1: Example of doubling inference problem. Genomes observed only for leaves (filled dots) of phylogeny. 2 = (diploid) unduplicated genome. Inferred doubling events indicated by red dots.

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. 4.1.

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.

4.3 Basic algorithms.

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 opera-

tions necessary to convert one genome X into another Y. 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. Genomic distance is defined only between undoubled genomes.

The essence of the various algorithms[1] for genomic distance resides in the "breakpoint graph" of the two genomes being compared. In this graph separate vertices are defined for the 5' and 3' end of each gene, and the adjacencies between two genes are represented by a black edge between the adjacent vertices in one genome and a gray edge between the adjacent vertices in the other genome. Though there are competing, and equivalent, formalisms for this, the graph thus defined decomposes into c cycles and paths of alternating gray and black edges, and it can be shown that the genomic distance is of the form $d(X, Y) = n + \chi - c$, where χ is the number of chromosomes in each genome.

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 $d(T, A \oplus A)$? Here we use a linear-time algorithm for solving this problem [51].

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-

[1]For our calculations in the present paper, we actually use the versatile rearrangement algorithm of Bergeron *et al.* [17], which we constrain to allow only the operations we have listed. This avoids difficulties of graph-theoretical programming associated with the breakpoint graph, while arriving at essentially the same result.

hard problem, we have a new implementation of a previous heuristic [1], which uses the principles of Bourque's MGR [26, 182], but is based on the constrained version of the Bergeron *et al.* [17] algorithm mentioned in the next to last paragraph. For the instances involving a few hundred genes we studied, this algorithm proposes a solution after a few minutes to a few hours computation on a MacBook computer.

4.4 Gene order phylogeny based on iterations of a median algorithm.

One of the main approaches to gene-order phylogeny is based on iterating the median algorithm to overlapping parts of a given binary branching tree as illustrated in Fig. 4.2. Though we can reconstruct the ploidy at the root using the methods of Section 4.2, and while knowing the gene order of the ancestor would, of course, be of great interest, the rearrangement approach produces inferences only about tree nodes of degree three or higher, and the reversibility of the rearrangement operations preclude any notion of the time direction of the tree edges. So whether the given tree is rooted or not is irrelevant to our analysis. In fact, we can assign the status of root to any node in the tree without changing the gene order results. Thus we work with unrooted trees, with the possibility that the position of the root is known from previous biological work or phylogenetic analysis.

An unrooted tree with all non-terminal vertices of degree three as on the left of the figure can be decomposed into overlapping median problems as on the right, where the three "given" genomes in the centre of the figure may be terminal vertices or ancestral vertices with previously calculated or assigned genomes.

The median algorithm is applied to each ancestral vertex in turn, perhaps in-

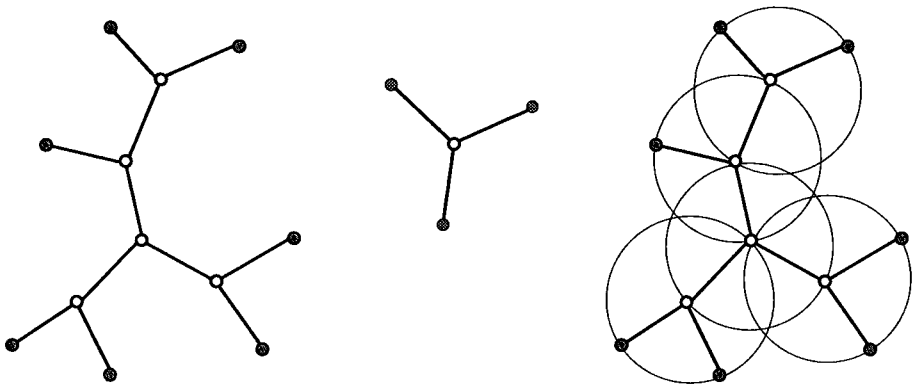

Figure 4.2: (left) Unrooted binary phylogenetic tree; filled dots represent given genomes, open dots represent ancestral genomes to be inferred. (centre) Smallest phylogeny requiring solution to the median problem. Shaded dots can represent either given or ancestral genomes. (right) Decomposition of phylogenetic inference problem on left into overlapping median problems.

volving several passes over the set of ancestors, until no further improvement is obtained. In the first pass, the ancestral vertices must be initialized, either using random genomes, copies of the closest given genome, or some other strategy.

4.5 Median-based phylogeny containing doubling descendants and undoubled genomes.

In the case where some of the given genomes are descended from whole genome duplication events, as evidenced by widespread pairs of paralogs, we can infer through the methods of Section 4.2 which ancestors are likely to have been doubling descendants. Then when we decompose the tree into median problems, we may encounter any of the four configurations in Fig. 4.3.

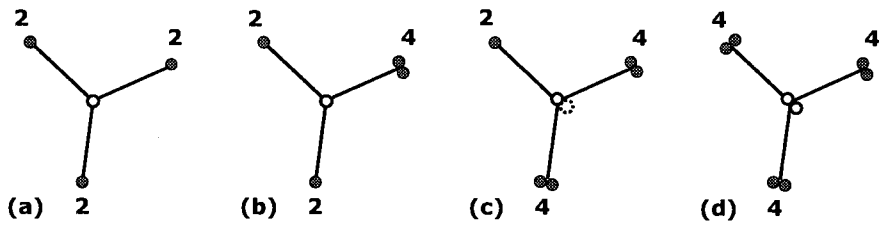

Figure 4.3: Four possible median configurations of undoubled (denoted "2") and doubling descendants (denoted "4"). (a) classical median problem with three undoubled genomes, (b) genome halving with two outgroups, (c) two doubling descendants from either one (include dotted circle) or two (without dotted circle) doubling events, as studied in this paper, (d) median problem with three doubling descendants. Although the inference of the median takes no account of the position of the root, the considerations of Section 4.2 imply that in all cases (a)–(d), we may assume either that the median is the root and the three edges lead to its offspring or that the upper left genome is the root or, especially in (d), on a path from the median to the root. In the latter case the median genome, which is unknown, is one of the two offspring of the upper left genome, the other two genomes shown are the offspring of the median, either or both of which can be given or unknown (but initialized).

4.5.1 The four cases.

We introduce some notation: 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. Each $g_{i,j}$ is signed positive or negative to indicate the reading direction of the DNA strand it is on. 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 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

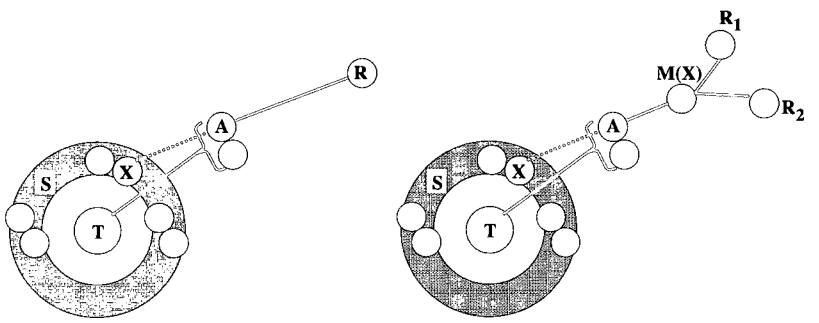

Figure 4.4: Halving a doubling descendent T, with one (R) or two (R_1, R_2) unduplicated outgroups. The double circles represent two copies of potential ancestral genomes, including solutions to the genome halving in the shaded ring $\mathbf{S}$. Ancestors labeled A, which minimize D, are found on best trajectories between solutions $X \in \mathbf{S}$ and the outgroup (or between X and M, the median of X, R_1 and R_2).

ancestor of T that incurred the doubling event.

4.5.2 Case (a).

This case is just the classical median problem, and may be solved using the previously cited methods.

4.5.3 Case (b).

We first consider a simpler problem called Genome halving with an outgroup. Consider T and and a related unduplicated genome R with markers orthologous to $g_1, \cdots, g_m$. The problem is to find an unduplicated genome A that minimizes

$$D(T, R) = d(R, A) + d(A \oplus A, T). \tag{4.5.1}$$

The solution in [238], as represented on the left of Fig. 4.4, 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.

We can use these same ideas for Case (b) in Fig. 4.3. 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

$$D(T, R_1, R_2) = d(R_1, M) + d(R_2, M) + d(A, M) + d(A \oplus A, T). \tag{4.5.2}$$

Our solution in [174], as on the right of Fig. 4.4, 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 $M(X)$.

4.5.4 Cases (c) and (d).

The present paper is most pertinent to Case (c). In the ensuing sections, although we will not directly address the case of two doubling descendants and one unduplicated genome, we will deal with the difficulties inherent in comparing two doubling descendants. The procedures we develop can then be extended to encompass the third genome, the unduplicated one, much as in Case (b).

Case (d), where the orthology relations among the two copies of a marker in one genome and the two copies in the other genomes is not known, remains to be investigated, although many of the considerations of Case (c) will be pertinent. If the various orthology relations are known, then Case (d) reduces to Case (a).

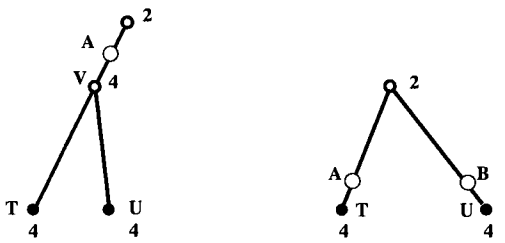

Figure 4.5: Left: Doubling, then speciation. Right: Speciation, then two independent doublings. Numbers indicate ploidy as in Figure 4.1.

4.6 The case of two doubling descendants.

Two related doubling 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. 4.5, 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 4.2 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 original halving algorithm.

4.6.1 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. There are two parts to algorithm, the first, **construct_SNGs**, being the partitioning of the adjacencies between markers in a doubling descendant T into a set of *supernatural*

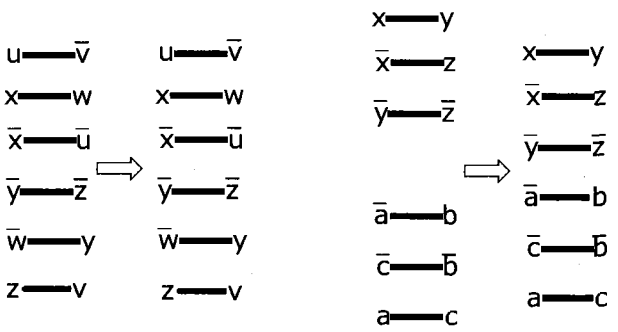

Figure 4.6: (left) Even-size natural graph completed by adding three pairs of gray edges in a way that maximizes the number of cycles. (right) Two odd-size natural graphs, containing x, y, z vertices and a, b, c vertices, respectively, combined into one supernatural graph so that three pairs of gray edges may be added.

graphs (SNGs) such that all edges in the breakpoint graph (defined in Section 4.3) induced by T and $A \oplus A$ connect vertices within the same SNG. The second part of the algorithm, **construct_**$A \oplus A$ is the actual construction of the adjacencies in $A \oplus A$ within the SNGs.

As a first step in **construct_SNGs**, whose pseudocode is presented below, each marker x in a doubling descendant is replaced by an oriented 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 [17]. For a vertex at the end of a chromosome, say y_t, it generates a edge of form $\{y_t, O\}$, where O is a dummy symbol.

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 $\{\bar{b}, d\}$. This key step is underlined in the pseudocode for the algorithm below.

If a natural graph has an even number of edges, as illustrated on the left of Figure 4.6, it can be shown that in all optimal ancestral doubled genomes, the edges coloured gray, representing adjacent vertices in the ancestor (recall the definitions in Section 4.3), and incident to one of the vertices in this natural graph, necessarily have as their other endpoint another vertex within the same natural graph.

Gray edges are added in pairs, so that each edge in the reconstructed doubled ancestor is duplicated, in accordance with the required output of the algorithm to produce two copies of the same ancestral genome. (Whether a is connected to b and $\bar{a}$ to $\bar{b}$ or a to $\bar{b}$ and $\bar{a}$ to b is immaterial, since the two versions of the same gene were originally identical.) These edges are added in such a way as to maximize the number of (alternating coloured) cycles and paths that make up this subgraph, in accordance with the definition of genomic distance in Section 4.3.

For natural graphs with an odd number of edges, it is impossible to complete them by adding pairs of gray edges. Nevertheless, as on the right of Figure 4.6, they may be grouped pairwise into *supernatural graphs* so that they may be completed with pairs of gray edges. Then, as with even natural graphs, 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.

The rules for drawing the gray edges, contained in **construct_**$A \oplus A$ (not repeated here), can be found in [51].

Along with the multiplicity of solutions caused by different possible constructions of supernatural graphs (choices of u and t to join), within these SNGs there may be many ways of drawing the gray edges during the procedure **construct_**$A \oplus A$. Without repeating here the lengthy details of **construct_**$A \oplus A$, it suffices to note that these alternate ways can be generated by choosing a different one of the vertices within the

Algorithm 1: construct_SNGs

Input: doubling descendant T
Output: set of SNGs for T
for *all* x *in* T **do**
define oriented pair of vertices (x_t, x_h) or (x_h, x_t), depending on whether X is signed positive or negative;
// The duplicate of a marker, e.g., $x = (x_t, x_h)$, is denoted $\bar{x} = (\bar{x_t}, \bar{x_h})$.
for *all pairs of neighbouring markers, e.g.,* (x_t, x_h) *and* (y_h, y_t), **do**
create unoriented black edge $\{x_h, y_h\}$ between adjacent vertices
for *each vertex at the end of a chromosome, e.g.,* y_t, **do**
create unoriented black edge $\{y_t, O\}$
All black edges and all vertices are initially unassigned to natural graphs;
while *there are still unassigned edges* **do**
choose one and assign it (and its two incident vertices) to a new natural graph s;
while *there is a vertex* a*, other than* O*, in* s *such that vertex* $\bar{a}$ *is not in* s **do**
assign $\bar{a}$ to s, as well as the previously constructed edge $\{\bar{a}, b\}$ and b
Natural graphs with even numbers of edges are automatically supernatural graphs;
while *there are natural graphs* u *and* t *both containing an odd number of edges* **do**
two such are joined to form a supernatural graph $s = u \cup t$;
// priority is given to constructing supernatural graphs with four O's instead of two. Other details on which u and t to join, and how to handle odd numbers of natural graphs, in [51]

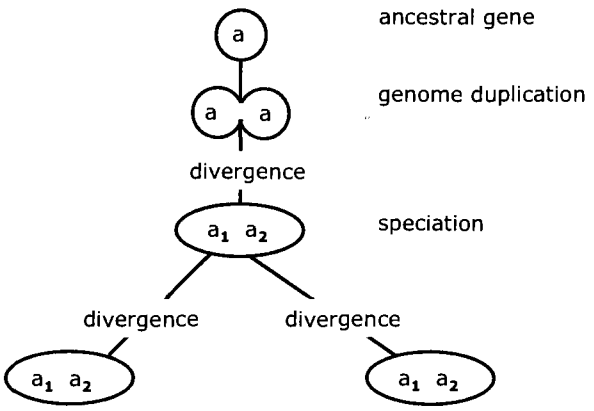

Figure 4.7: Why ortholog identification is pertinent to the doubling first scenario. The genes labeled a_1 in the two genomes are orthologous, as are the two labeled a_2.

SNG as a starting point.

4.6.2 Ortholog identification.

When comparing a doubling descendant to an undoubled related genome, there is no question of which of the two paralogs of a gene in the former is the true ortholog of the corresponding single gene in the latter: the two paralogs have precisely the same historical relationship with the single gene because they were both identical at the instant of doubling.

When comparing two descendants of the same doubling event that have subsequently diverged though a speciation event, however, we cannot avoid the problem of identifying two orthologous pairs of genes, each involving one gene in each genome. Evolution will have affected the two copies of the original gene differently in the time interval between the doubling and speciation events, for example with respect to the genomic contexts they are found, so that at the time of speciation and thereafter, there are two distinct pairs of orthologous genes deriving from a single gene pre-duplication, as in Fig. 4.7. Depending on the time elapsed between doubling and

speciation, sequence divergence analysis may help identify the correct orthologies via gene tree/species tree methodology, and further phylogenetic information may also be brought to bear, but this type of information may not always be available.

Another way of deducing the orthologies is through parsimony. Incorrectly posited orthologies will tend to increase the genomic distance between the two doubling descendants, so it is reasonable to look for the ensemble of orthology assignments that minimizes this distance.

In Sections 4.6.3 and 4.8, we will deal with both the case where the correct orthologies are given, and the case where they must be inferred. Note that if the two doubling descendants are descended from separate, post-speciation, doubling events, as to be described in Section 4.6.4 and also in Section 4.8, the question of correct orthology identification does not arise, much as in the case of one doubling descendant and one unduplicated genome, because the two paralogs in one doubling descendant are both related in exactly the same way to each of the two paralogs in the other; each pair is derived independently from the same single-copy gene.

Note that when we carry out the halving algorithm, this effectively partitions the genes among the chromosomes, and tells us which copy of each gene were together on a chromosome in the original doubled genome (original synteny). Of course, the other copy of each of these genes must also all have been together on a single chromosome. The algorithm, however, does not partition the ancestral chromosomes thus defined into two classes corresponding to the two original copies of the doubling genome, as this distinction has no biological meaning past the moment of doubling, and so has no trace in the doubling descendant. In any case, given the considerable degree of non-uniqueness in the halving algorithm, inferences of original synteny based on halving alone may be unreliable.

4.6.3 Doubling first.

Given two doubling descendants T and U as on the left of Fig. 4.5, 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 4.3) 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 T and U themselves, where each successive pair of genomes along the trajectory differ by exactly one rearrangement operation.

Before explaining a greedy (with look-ahead) heuristic for a solution to the constrained version of the problem, we recall the edge notation we use to represent the adjacencies in a genome [17]. 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 initially define $T^*=T, U^*=U$ and associate this pair with the root node of a search tree. 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 will measure whether

the operation actually diminishes $d(V, A \oplus A)$, while the x will measure whether the operation increases the potential of diminishing $d(V, A \oplus A)$ in a subsequent operation.

In a greedy approach, y should be more heavily weighted than x. But how do we justify the coefficient 6 of y? 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 $d(V_2, U) = d(V_1, U) - 1$. Let A_1 and A_2 be solutions of the halving problem for V_1 and V_2, respectively. Even if $A_2 \neq A_1$, we have $|d(V_1, A_1 \oplus A_1) - d(V_2, A_2 \oplus A_2)| \leq 1$, because only one operation is involved. We set $y = d(V_1, A_1 \oplus A_1) - d(V_2, A_2 \oplus A_2) + 1$, so that y is in the range $[0, 2]$.

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 $x = h + 1$. Then x is in the range $[1, 6]$.

The score W is in the range $[1, 18]$. Note that the effect of y always outweighs that of x in in the formula $W = x + 6y$. Even if $x = 6$, this can only happen if $y = 2$, since having all five advantageous possibilities can only happen in the context of V_2 having a smaller halving cost than V_1. Thus the greedy approach mandates a coefficient of y of at least 6; we chose a coefficient of exactly 6 in order to maximize the effect of the look-ahead.

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. Pseudocode for this algorithm, **search_trajectory**, follows. Because of the asymmetric way it treats T and U, we repeat it reversing the roles of the two genomes, and choose the better solution of the two.

Algorithm 2: search_trajectory

Input: two doubling descendants T and U with given orthologies between each pair of paralogs in T and corresponding pair of paralogs in U
Output: doubling descendant V on trajectory between T and U with minimum halving cost
define the pair $T^* = T, U^* = U$ to constitute the root node of a search tree;
while $T^* \neq U^*$ **do**
  for *each possible operation decreasing* $d(T^*, U^*)$ **do**
    calculate y, x and hence W_T and W_U according to the actual and potential changes to the halving distance
  if *the highest scoring operation has* $W_T \geq W_U$ *and* $W_T \geq 6$ **then**
    apply this operation to update T^* and create a new node (T^*, U^*) on the search tree
  if *the highest scoring operation does not satisfy the above conditions on* W_T, *but* $W_U > 1$ **then**
    apply this operation to update U^* and create a new node (T^*, U^*) on the search tree
  if *no operation satisfies the conditions on* W_T *or* W_U **then**
    resume search at the previous node on the search tree
Find the optimal $V = T^*$ or $V = U^*$ on the search tree

This search requires $O(m^3)$ time in the worst case for a depth-first traversal of the search tree, where m is the number of markers. This follows since the traversal, which provides an initial estimate of the solution and an initial upper bound on the

solution cost, verifies the **while** conditions for $O(m)$ pairs (T^*, U^*) and examines $O(m^2)$ possible operations each time. The total time to find the locally optimal solution depends on how large a search tree we are willing to maintain.

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%.

4.6.4 Speciation first.

In Section 4.6.3, $d(T, V) + d(V, U)$ 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 **halve_two** simultaneously halves T and U, with a subroutine **add_gray_edges** 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. The SNGs have previously been ordered by a subroutine **sort_SNGs** to favour this maximization.

This search also requires $O(m^3)$ time in the worst case for a depth-first traversal of the search tree, where m is the number of markers, since the number of SNGs to construct is $O(m)$ and both **sort_SNGs** and **add_gray_edges** require $O(m^2)$ to compare SNGs from T with those from U. This traversal provides an initial estimate of the solution and an initial upper bound on the solution cost. The total time to find the locally optimal solution depends on how large a search tree we are willing to

maintain.

Algorithm 3: sort_SNGs

Input: σ_T and σ_U, the set of supernatural graphs for T and U, respectively
Output: ordered sets of SNGs $\sigma_T^{(1)}$ and $\sigma_U^{(1)}$
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 $\sigma_U^{(0)} = \sigma_U \setminus \sigma_U^{(1)}$;
`// if there are no SNGs with only two black edges, some other SNG is chosen to initialize either` $\sigma_T^{(1)}$ `or` $\sigma_U^{(1)}$
while *there remain SNGs in* $\sigma_T^{(0)}$ *or SNGs in* $\sigma_U^{(0)}$ **do**
  while *there remain SNGs in* $\sigma_T^{(0)}$ *and either* $\sigma_U^{(0)}$ *is empty or the number of black edges in* $\sigma_T^{(1)} \leq$ *the number in* $\sigma_U^{(1)}$ **do**
    `// find a SNG in` $\sigma_T^{(0)}$ `to move to` $\sigma_T^{(1)}$
    for *each SNG* s *in* $\sigma_T^{(0)}$ **do**
      `// calculate the largest number of gray edges it could have in common with SNGs in` $\sigma_U^{(1)}$
      for $i = 1, \cdots, |\sigma_U^{(1)}|$ **do**
        Let k_i be the number of vertices SNG s has in common with u_i, the $i-$th SNG in $\sigma_U^{(1)}$;
        the number of gray edges they could have in common is $\leq \lfloor \frac{k_i}{2} \rfloor$
      then the score of s is $\sum_{i,\cdots,|\sigma_U^{(1)}|} \lfloor \frac{k_i}{2} \rfloor$
    add the highest scoring s to $\sigma_T^{(1)}$
  while *there remain SNGs in* $\sigma_U^{(0)}$, *and either* $\sigma_T^{(0)}$ *is empty or the number of black edges in* $\sigma_U^{(1)} <$ *the number in* $\sigma_T^{(1)}$ **do**
    find a SNG in $\sigma_U^{(0)}$ to move to $\sigma_U^{(1)}$ in the analogous way as for T

4.7 Simulations

For simulations of the doubling first model, we chose the following parameters: five chromosomes, number of markers $m = 200$, inversions to translocations proportion 5:3 or 10:1, random choice of chromosomes to be rearranged, random breakpoints on chromosomes, true orthology relations available for reconstruction. Our algorithm accurately reconstructs the number ν of rearrangements (ten replications for each

Algorithm 4: add_gray_edges

Input: ordered sets of SNGs $\sigma_T^{(1)}$ and $\sigma_U^{(1)}$

Output: search tree where each node is the set of SNGs of T and U with partial assignments of gray edges

At the root of the search tree, add gray edges to all 2-edge SNGs in σ_T and σ_U;

`// if there are no SNGs with only two black edges, some other SNG is chosen`

while *there remain SNGs in* σ_T *or* σ_U *without gray edges* **do**

  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* $\sigma_T \leq$ *the number of gray edges in* σ_U **do**

    let s be the first SNG in σ_T (using the order in $\sigma_T^{(1)}$) that has no gray edges, where s has v_s vertices **for** $i = 1, \ldots, v_s$ **do**

      starting with the i-th vertex, add gray edges to s according to the deterministic procedure in the halving algorithm, and retain a new node in the search tree if starting with this i-th vertex maximizes the number of gray edges in common with σ_U

    `// At this point we can disregard orthology assignments in counting potential gray edges, since these may be reassigned with no other consequence`

  while *there remain SNGs in* σ_U *without gray edges and either all SNGs in* σ_T *have gray edges or the number of gray edges in* $\sigma_U <$ *the number of gray edges in* σ_T **do**

    add gray edges to the first SNG in σ_U that has no gray edges using the same procedure as with T, thus creating one or more new nodes in the search tree

Algorithm 5: halve_two

Input: T and U

Output: (optimal) genome halvings A and B, of T and U respectively, which minimize $d(A, B)$

construct_SNG for T and for U;

sort_SNGs;

add_gray_edges;

find closest halvings A and B on search tree

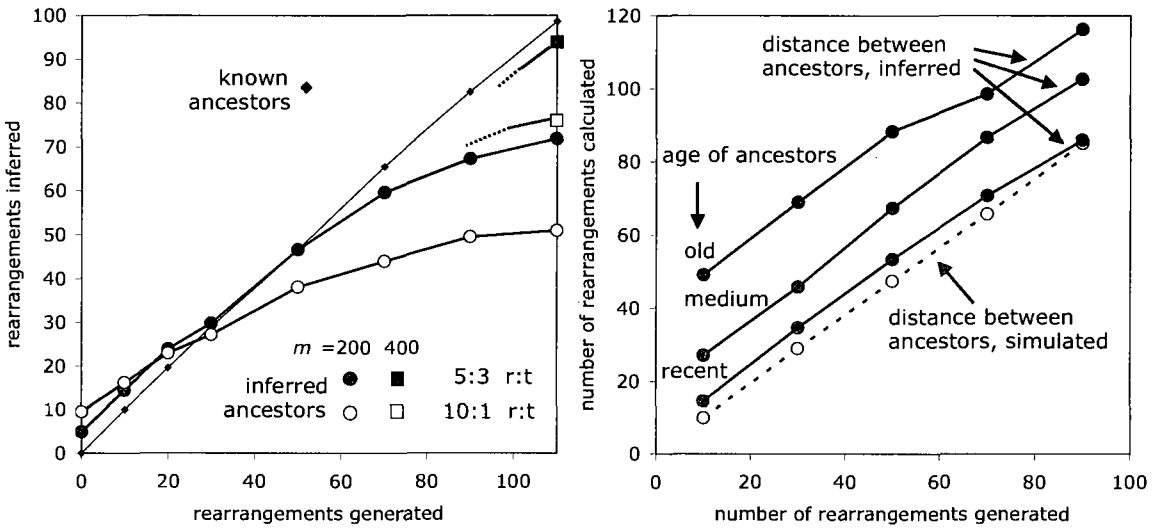

Figure 4.8: Estimated distance: left, between doubling and speciation (age of ancestor=50) in doubling first model, right, between unduplicated ancestors (ages: old=80, medium=50, young=20) in speciation first model.

value of ν) between the doubling event and the speciation event, as long as this is not too large (Fig. 4.8, left), about $\nu = 50$ for a reversals to translocations ratio of 5:3 and $\nu = 30$ for a ratio of 10:1. With a longer interval between doubling and speciation, the halving algorithm reconstructs the unduplicated ancestor too economically. The shortfall is a function not only of the number ν of rearrangements in the simulation, but also of the proportion of reversals versus translocations ($r : t$) and of the number of markers m. If the number of markers is doubled from 200 to 400, the shortfall in the inferred number of rearrangements is largely corrected, as indicated by the square dots in the figure.

Thus in any inference based on the doubling first model, we must account for the fact that small m, large ν and high inversion proportion may bias the results towards an underestimate of the cost of fitting the model.

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 doubling descendant in the simulation (data not shown here), the estimated unduplicated ancestors are further apart than the genomes actually generated in the simulation (Fig. 4.8, right). 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.

In any use of the speciation-first model, we must account for the fact that a large number of rearrangements in the doubled genomes may bias inter-ancestor distance towards an overestimate of the cost of fitting the model.

Nonetheless, the superior accuracy and efficiency 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. 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. Running our speciation first algorithms to completion on the two doubling descendants, we reached an inter-ancestor distance $d(A, B) = 84$ (instead of the simulated distance of 75) after three

experiment	simulated			sample			**halve_two**	
number	d	t	$\hat{d}$	size	$\hat{d}$	time	$\hat{d}$	time
1	$45+38$	$42+50$	75	1.0×10^6	87	24h	84	3.0h
2	$20+20$	$60+60$	38	1.47×10^6	70	35h	57	0.8h
3	$40+40$	$40+40$	71	0.22×10^6	83	12h	76	1.5h
4	$60+60$	$20+20$	100	0.26×10^6	101	17h	100	0.6h

Table 4.1: Comparison of performance of the halve_two algorithms with the sampling method,varying divergence d between the ancestors and subsequent rearrangement distance of the doubled genomes t.

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. We stopped the latter calculation when several hours elapsed without any improvement. These results appear as experiment number 1 in Table 4.1. This table also contains the results from three similar experiments with different values for divergence d between the ancestors and subsequent evolution of the doubled genomes t.

While these experiments confirm that there is still a long way to go to eliminate the overestimate bias in inferring d, they indicate that with a tiny fraction of the computing time, **halve_two** achieves substantial reductions in this bias.

4.8 Genome doubling in yeast

Wolfe and Shields [227] discovered an ancient genome doubling in the ancestry of *Saccharomyces cerevisiae* in 1997 after this organism became the first to have its genome sequenced [62]. According to [95], the recently sequenced *Candida glabrata* [46] shares this doubled ancestor. We extracted data from YGOB (Yeast Genome Browser) [31], on the orders and orientation of the exactly 600 genes identified as

duplicates in both genomes, i.e., 300 duplicated genes.

Orthology considerations. The algorithm in Section 4.6.3 requires input about orthology relations among the genes in the two doubling descendants. This question is not pertinent to the algorithm in Section 4.6.4. We were able to obtain this information, i.e., about which of the two duplicates in one genome is orthologous to which duplicate in the other genome, from YGOB.

We also calculated a separate estimate of the orthology relations, motivated by the parsimony argument in Section 4.6.2, using the following heuristic:

Algorithm 6: find_orthologies

Input: two doubling descendants T and U, genome size $2m$, block size parameter k

Output: two resolved orthologies (inferred) between each pair of paralogs in T and corresponding pair of paralogs in U

for *each paralogous marker* **do**
  randomly construct two orthology relations each connecting one of the paralogs in T to one of the paralogs in U

calculate $d = d(T, U)$;

construct a list of all $2m$ markers by concatenating all chromosomes of genome T. This list can be grouped into $2m - k + 1$ blocks of k consecutive markers;

for *all blocks of k consecutive markers* **do**
  try changing all the orthology assignments in the block of k genes;
  calculate $d' = d(T, U)$ based on this trial reassignment;
  if $d' < d$ **then**
    retain the changed assignments and set $d' = d$

Continue cycling through the list until no further decrease is possible

With the yeast data, increasing k decreased $d(T, U)$, though the improvement after $k = 5$ was minimal. For random orthology assignments, $d(T, U) = 463 \pm 7.45$, while for $kmax = 5$, the algorithm produces $d(T, U) = 327.3 \pm 4.16$. For comparison, the YGOB assignment, in which we have the most confidence, gives $d(T, U) = 337$. If we initialize the algorithm with this assignment, it is improved to $d(T, U) = 318$. We

use the assignment giving the latter value as our estimate of the most parsimonious assignment.

Competing hypotheses. Though the results of the algorithm in Section 4.2 suggests that the theory in [95] 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. We will compare the analysis of the yeast data, using the method in Section 4.6.3, assuming the doubling first model depicted on the left of Fig. 4.5, with the analysis in Section 4.6.4, assuming the speciation first model depicted on the right in Fig. 4.5.

The key point to keep in mind in comparing analyses based on these two different models is that they are both liable to considerable and variable bias, and we must avoid choosing one model simply because its results tend to be biased towards smaller distances while the other is biased towards larger values. The way we do this is first to analyze the yeast data by both of the two methods. Then for each model, we use the inferred rearrangement and halving distances to simulate new data sets with the same number of markers. Finally, we apply both methods to both simulated data sets. Only if each method is superior at inferring the data generated by its own model can we have some confidence that the method that most economically analyzes the yeast data reflects the appropriate model. Otherwise, if one method systematically produced better analyses, even for data generated by the other model, our inferences about the yeast data would be spurious.

Thus, we analyzed the yeast data using the doubling first and speciation first algorithms. The results appear in the centre row of Table 4.2. (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 (**search_trajectory**) and speciation first **halve_two**), to both sets of data. Note first in Table 4.2 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.

Table 4.2: Doubling first (d.f) and speciation first (s.f.) analyses each produce a more parsimonious analysis (asterisked) 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
Control								
sim by d.f.:	102	213	166	481				
inferred:	119	181	157	**457***	214	163	255	**632**
yeast:	92	245	168	**505***	193	182	250	**625**
	122	215	184	**521***				
Control								
sim by s.f.:					177	164	225	566
inferred:	146	354	104	**604**	164	228	197	**589***

The rows in Table 4.2 show that the doubling first analysis is better (indicated by an asterisk) 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 simulated controls assure that the biases in the two methods cannot be invoked, so our analysis confirms the hypothesis in [95].

Note that the way we have controlled our analysis of the yeast data by using simulations also eliminates the effects of any general bias of rearrangement methods on the results (though no such bias shows up in comparing the number of events generated in the simulations to the number inferred, across the six sets of simulations in Table 4.2). The key point is that the speciation first analysis gives better results for the data generated by speciation first, and that the doubling first analysis gives far better results for the data generated by doubling first. This validates our use of the methods as diagnostic of evolutionary history. Thus the fact hat the doubling first analysis gives far better results than the speciation first analysis when applied to the yeast data makes it very unlikely that these data were generated by speciation first.

The doubling first analysis of yeast in Table 4.2 used the YGOB orthology assignment, and for comparability's sake the doubling first analysis of the simulated data used an orthology assignment traced through the actual generation of the data. When we use instead **find_orthologies**, the parsimonious orthology assignment, the totals of 505 and 521 change to 491 and 501, respectively. This reflects the tendencies of parsimony methods to underestimate rearrangement distances (by 3 or 4% in this case), but does not affect our conclusions.

4.9 Conclusions

We return to the question of a median between an unduplicated genome and two doubling descendants raised in Section 4.5.4. Though we have focused in this paper on the problems of comparing only the two doubling descendants, both of our algorithms can be directly integrated with existing techniques to take care of the remaining genome. The doubling first algorithm identifies a doubling descendant whose halving

distance is minimal. This genome can be used directly as input to the genome halving with an outgroup algorithm [238] discussed at the beginning of Section 4.5.3, though with the reservations mentioned below. This solves the version of Case (c) in Fig. 4.3 where the median is a doubling descendant.

The version of Case (c) where the median is an unduplicated genome cannot be so directly addressed using the methods developed here. Nevertheless, the basic principle of the speciation first algorithm, namely the search for common gray edges, can easily accommodate another criterion, namely the prioritization of gray edges representing adjacencies in the third (unduplicated) genome of the median problem. This is the subject of promising ongoing research.

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, via the median problem as mentioned above.

Our focus in this work has been on estimating the chromosomal structure of the doubled and undoubled ancestors of given genomes, in a given phylogeny. Although

we are primarily interested in working out the mathematical aspects of these questions on the basis of gene order data, it is clear that information at the level of DNA or amino acid sequence could also be helpful. For example, under the speciation first model, the two copies of a gene within each species should be more similar than the between-gene comparisons, while under the doubling first model, each copy of a gene should have its own unique ortholog in the other species, more similar than its own paralog. We should also note that although we work with phylogenies, these are given and not to be inferred. We are interested here in the small phylogeny problem and not the large problem of actually inferring phylogenies. And though questions of cross-validation of sequence-based and order-based phylogenetics, or of their preferred domains of application, are of great interest, they are outside the scope of the present work, which aims only at inferring gene order, and through that, the occurrence of evolutionary events that affect gene order.

Chapter 5

Guided genome halving: hardness, heuristics and the history of the Hemiascomycetes

Chunfang Zheng, Qian Zhu, Zaky Adam, David Sankoff. 2008. Guided genome halving: hardness, heuristics and the history of the Hemiascomycetes. Bioinformatics 24: i96–i104.

I was responsible for designing the algorithm and getting all of the results in this work. I also collaborated fully in writing and preparing the manuscript, especially technical write-ups and graphics. Qian Zhu extracted the yeast data from YGOB [31] and applied the 3D principal coordinates analysis to the matrix of genomic distances. Zaky Adam provided helpful discussion and advice about the median analysis.

Abstract

Motivation:

Some present-day species have incurred a whole genome doubling event in their evolutionary history, and this is reflected today in patterns of duplicated segments scattered throughout their chromosomes. These duplications may be used as data to "halve" the genome, i.e., to reconstruct the ancestral genome at the moment of doubling, but the solution is often highly non-unique. To resolve this problem, we take account of outgroups, external reference genomes, to guide and narrow down the search.

Results:

We improve on a previous, computationally costly, "brute force" method by adapting the genome halving algorithm of El-Mabrouk and Sankoff so that it rapidly and accurately constructs an ancestor close the outgroups, prior to a local optimization heuristic. We apply this to reconstruct the pre-doubling ancestor of *S. cerevisiae* and *C. glabrata*, guided by the genomes of three other yeasts that diverged before the genome doubling event. We analyze the results in terms 1) of the minimum evolution criterion, 2) how close the genome halving result is to the final (local) minimum, and 3) how close the final result is to an ancestor manually constructed by an expert with access to additional information. We also visualize the set of reconstructed ancestors using classic multidimensional scaling to see what aspects of the two doubled and three unduplicated genomes influence the differences among the reconstructions.

5.1 Introduction

Whole genome doubling (WGD) is a rare but important type of evolutionary event, often giving rise to major new lineages. In its various forms it has occurred across the eukaryotic spectrum, from the pathogenic protist *Giardia* to the ancestor of brewer's yeast, to most of the plant lineages, to several insect species, to the salmonid fishes, to amphibians and even to mammalian species.

WGD is followed, over evolutionary time, by genome rearrangement through intra- and interchromosomal movement of genetic material. The phylogenetic study of synteny, gene order and chromosomal evolution becomes blocked because of the extraordinarily high rates of paralogy in the species descended from the WGD compared to sister species that diverged before the WGD. If we could infer the ancestral genome that underwent the WGD, this difficulty would be resolved. Thus the genome halving problem is to reconstruct the ancestral genome on the basis of a decomposition of the present-day genome into a set of apparently duplicated blocks of genes or DNA sequence dispersed among the chromosomes[1]. A linear-time algorithm to find the ancestral genome that minimizes the genomic distance to the present-day genome has been available for some time [50, 51]. Unfortunately, the solution to the combinatorial optimization problem is not always directly interpretable as a solution to the evolutionary biology problem. First, the algorithmic result suffers from severe non-uniqueness. Second, in common with most methods of inferring history, we have no direct way to verify if the answer is correct. Our goal is to counteract these problems, first by guiding the reconstruction by one or more reference, or outgroup, genomes and second, by checking our results for a particular data set against an ancestor genome manually reconstructed by an expert.

[1]Sequence analysis tools for dating duplication events are not pertinent to this problem since all pairs of duplicates in the doubled genome were generated at the same historical moment

If our guided reconstruction method were to be feasible and accurate it could have wide application. One or more descendants of a WGD event co-occur with unduplicated sister species in many phylogenies. This is most prevalent among plants where, for example, the poplars and willows descend from a common WGD, while the closely related eurosid angiosperms like papaya diverged before this event, but it also occurs in yeast, where brewer's yeast and several sister species share an origin in an ancestral WGD, while other closely related species have earlier divergence dates, in fish, where the salmonid species like trout and salmon originate in a WGD event after diverging from the related osmerid fish, in mammals, where some genera of viscacha rodents share a WGD history while their relationship with very similar octodontids predates this. In protists, the important pathogen genus *Giardia* has undergone a form of WGD, while the related enteromonad parasites have not, though this may be due to a post-WGD loss rather than an early divergence. This very partial list of examples emphasizes species whose genomes have been sequenced or for which serious sequencing projects are underway or are being actively promoted.

We first explored the idea of guided reconstruction for the ancestral doubled genome of the maize (*Zea mays*) genome, with the rice (*Oryza sativa*) and sorghum (*Sorghum bicolor*) genomes[2] as outgroups [238]. Our strategy was to generate all the 1.5×10^6 solutions to the genome halving problem for the maize genome, and to identify the subset, containing 10-20 solutions that have a minimum rearrangement distance with the rice (or sorghum) genome. We followed this with a local improvement heuristic searching outside the immediate set of optimal halving solutions to find the genome A that minimizes the sum of the distance between the doubled form $A \oplus A$ and present-day maize plus the distance between rice (or sorghum) and the

[2]All cereals underwent earlier WGD event(s), but the effects of these can be filtered out on the basis of greater sequence divergence.

pre-doubling form A.

While this approach was feasible with the 34 doubled blocks in maize, present in one copy in each outgroup, the heuristic search step was time-consuming, given that the starting points were relatively far from optimal. Then we attempted to reconstruct the ancient doubled yeast genome from which *Saccharomyces cerevisiae* is descended, guided simultaneously by both of the undoubled outgroup genomes *Ashbya gossypii* and *Kluyveromyces waltii* [174]. In these data the number of doubled genes we used was an order of magnitude greater than the number of blocks in the cereals data, and the number of solutions to the halving problem astronomical. It is not feasible to exhaustively search the halving solutions to find those that are closest to the outgroups, to say nothing of the heuristic search step. Instead we tried working with a sample of halving solutions, hoping to generate at least one initialization leading to a good solution. It was not clear, however, how large the sample should be, or how to validate the results, since the local optima found in that study remained fairly far apart, as measured by genomic distance.

The facts that halving guided by a single outgroup involves only two genomes, and that both of its component parts, halving and distance calculation, are basically linear-time, suggests that this problem might be susceptible to a polynomial-time analysis, in contrast to problems such as the "median problem" for three or more genomes, which are NP-hard [35, 46, 153]. We dispose of this hope at the outset, by showing that the simplest problem of halving guided by one outgroup is NP-hard.

Nevertheless, in the ensuing sections, we seek to replace the "brute force" approach of generating unconstrained halving solutions first, i.e., before taking into consideration the outgroup genome(s). Instead, we inject all pertinent information derivable from the outgroup(s) into the halving algorithm, influencing hitherto arbitrary choices in that algorithm so that the halving solution is guided towards the

outgroup(s).

We analyze data on two yeasts descended from the same doubling event, *Saccharomyces cerevisiae* and *Candida glabrata*, to try to reconstruct the original doubled genome. Three related outgroup species are currently available in the Yeast Gene Order Browser (YGOB) [31]: *Ashbya gossypii, Kluyveromyces waltii* and *Kluyveromyces lactis*. YGOB also furnishes an estimate of the ancestral doubled genome painstakingly reconstructed by Jonathan Gordon on the basis of multiple sources of information.

Our new algorithm greatly improves the accuracy of our results, while drastically reducing the computational effort, both in generating halving solutions and in the local optimization search. We compare this new approach to the sampling approach, with and without the local optimization step, from the viewpoints of the objective function value obtained and computing time. We apply our method to all combinations of the two descendants of the doubled ancestor and four single genomes, the three species already mentioned plus Gordon's manually reconstructed ancestor.

We also use data-analytic methods to compare the our inferred pre-doubling genomes to each other and to the Gordon construct.

5.2 Problem statement.

Although the idea of guided genome halving is not difficult, the prerequisite for understanding the analysis is some knowledge of standard genome rearrangement problems, namely genomic distance, genome halving and genome median. We can only sketch these in Sections 5.2.2, 5.2.3 and 5.2.4 before enunciating the GGH problems in Section 5.2.5. In Section 5.3, we discuss the algorithms for these problems.

5.2.1 Genomes and rearrangement operations.

A genome G is represented by a set of strings (called *chromosomes*) of form $\{g_{11}\cdots g_{1n_1},...,g_{\chi 1}\cdots g_{\chi n_\chi}\}$, where $n = n_1 + \cdots + n_\chi$ and $\{|g_{..}|\} = \{1,\cdots,n\}$; i.e., each integer $i \in \{1,\cdots,n\}$, representing a gene or other marker, appears exactly once in the genome and may have either positive or negative polarity. The biologically-motivated operations generally include[3] inversions (implying as well change of sign, i.e., change of strand) of chromosomal segments, e.g., $h_1\cdots h_u\cdots h_v\cdots h_m \rightarrow h_1\cdots -h_v\cdots -h_u\cdots h_m$, and reciprocal translocations, e.g., $h_1\cdots h_u\cdots h_l, k_1\cdots k_v\cdots k_m \rightarrow h_1\cdots k_v\cdots k_m, k_1\cdots h_u\cdots h_l$.

5.2.2 Genomic distance.

The genome rearrangement distance $d(G,H)$ is defined to be the minimum number of operations necessary to convert one genome G into another H.

The breakpoint distance. We say there is a "shared adjacency" if the signed integer $g_{i,j+1}$ immediately follows $g_{i,j}$ on a chromosome in H as well as on the i-th chromosome in G, or if $-g_{ij}$ follows $-g_{i,j+1}$ in H. There are also shared adjacencies if g_{i1} or $-g_{in_i}$ are first terms on chromosomes in H or if g_{in_i} or $-g_{i1}$ are last terms on chromosomes in H. Then if G and H have the same number of chromosomes χ, the breakpoint distance $d_B(G,H)$ is defined to be $n+\chi-$ the number of shared adjacencies.

[3]See [231] for a more general inventory.

5.2.3 Genome halving.

Let T be a genome consisting of ψ chromosomes and $2n$ genes $a_1^{(1)} \cdots, a_n^{(1)}; a_1^{(2)}, \cdots, a_n^{(2)}$, dispersed in any order on the chromosomes. For each i, we call $a_i^{(1)}$ and $a_i^{(2)}$ "duplicates", but there is no particular property distinguishing all elements of the set of $a_i^{(1)}$ in common from all those in the set of $a_i^{(2)}$. A potential "doubled ancestor" of T is written $A' \oplus A''$, and consists of 2χ chromosomes, where some half (χ) of the chromosomes, symbolized by the A', contains exactly one of $a_i^{(1)}$ or $a_i^{(2)}$ for each $i = 1, \cdots, n$. The remaining χ chromosomes, symbolized by the A'', are each identical to one in the first half, in that where $a_i^{(1)}$ appears on a chromosome in the A', $a_i^{(2)}$ appears on the corresponding chromosome in A'', and where $a_i^{(2)}$ appears in A', $a_i^{(1)}$ appears in A''. We define A to be either of the two halves of $A' \oplus A''$, where the superscript (1) or (2) is suppressed from each $a_i^{(1)}$ or $a_i^{(2)}$. These χ chromosomes, and the n genes they contain, $a_1, \cdots, a_n$ constitute a potential "doubled ancestor" of T.

The genome halving problem for T is to find an A for which some $d(A' \oplus A'', T)$ is minimal.

5.2.4 The median problem.

Let P, Q and R be three genomes on the same set of n genes. *The rearrangement median problem is to find a genome M such that $d(P, M) + d(Q, M) + d(R, M)$ is minimal. The breakpoint median problem is to find a genome M such that $d_B(P, M) + d_B(Q, M) + d_B(R, M)$ is minimal.*

5.2.5 Guided genome halving.

As in Section 5.2.3, let T be genome consisting of ψ chromosomes and $2n$ genes $a_1^{(1)} \cdots, a_n^{(1)}; a_1^{(2)}, \cdots, a_n^{(2)}$, dispersed in any order on the chromosomes, where for each i, genes $a_i^{(1)}$ and $a_i^{(2)}$ are duplicates. Any genome R is a reference or outgroup genome for T if it contains the n genes $a_1, \cdots, a_n$.

There are a number of different formulations possible for guided genome halving (GGH), depending on the genomic distance used, and the number of outgroups doing the guiding. Here we will study the cases of one outgroup [238] and two outgroups [174], using the genomic distance d defined in Section 5.2.2, and we will also analyze the complexity of the one outgroup problem in the context of the breakpoint distance d_B.

Let R be a reference genome for T. The GGH problem with one outgroup is to find (an estimated ancestral) genome A such that some $d(R, A) + d(A' \oplus A'', T)$ is minimal. Let R_1 and R_2 be two reference genomes for T. The GGH problem with two outgroups is to find A and a median genome M such that some $d(R_1, M) + d(R_2, M) + d(A, M) + d(A' \oplus A'', T)$ is minimal.

5.3 Algorithms for genome distance, genome halving and the genome median.

5.3.1 Distance.

Rearrangement algorithms [210] can be formulated in terms of the bi-coloured "breakpoint graph", where each end (either 5′ or 3′) of a gene in genome G is represented by a vertex joined by a black edge to the vertex for adjoining end of the adjacent

gene, and these same ends, represented by the same $2n$ vertices in the graph, are joined by gray edges determined by the adjacencies in genome H. In addition, if G has χ chromosomes, assuming without loss of generality that this is at least as many as H, each vertex representing a first or last term of some chromosome in G only is connected by a black edge to an individual "cap", or dummy, vertex so that there are $2n + 2\chi$ vertices in all. The breakpoint graphs necessarily consist of disjoint alternating colour cycles and/or paths, and it can be shown that, with some rare, easily identifiable exceptions, $d(G, H) = n + \chi - c - \Pi$, where c is the number of cycles and Π the number of paths terminating in at least one cap. Calculating the distance can be done in time linear in n.

The actual operations, $d(G, H)$ in number, may be reconstructed by successively choosing certain large cycles and paths in the breakpoint graph to split into two, corresponding to a reversal or translocation, until there are $n - \chi$ cycles each made up of two vertices, a black edge and a gray edge, and 2χ paths each containing one cap and one chromosome-terminating gene vertex connected by a black edge. This requires somewhat more than linear time.

The breakpoint distance d_B is easily calculated by storing all adjacencies of G as it is input, and verifying for each g_{ij} as it is encountered when H is input, whether its successor is $g_{i,j+1}$.

5.3.2 Halving.

In the rearrangement distance algorithm, construction of the breakpoint graph is an easy step. The genome halving algorithms [51] also make use of the breakpoint graph, but the problem here is the more difficult one of building the breakpoint graph where one of the genomes (the doubled ancestor $A' \oplus A''$) is unknown. This is done by

segregating the vertices of the graph in a natural way into subsets, such that the vertices of all cycles must fall within a single subset, and then constructing these cycles in an optimal way within each subset so that the black edges correspond to the structure of the known genome T and the gray edges define the adjacencies of $A' \oplus A''$.

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

Following this, for each pair of neighbouring genes, say (a_t, a_h) and (b_h, b_t), the two adjacent vertices a_h and b_h are linked by a black edge, denoted $\{a_h, b_h\}$ in the notation of [17]. For a vertex at the end of a chromosome, say b_t, it generates a virtual edge of form $\{b_t, \text{end}\}$. Note that the use of "end" instead of "cap" reflects a somewhat different bookkeeping for the beginnings and ends of chromosome in the halving algorithm compared to the distance algorithm in Section 5.3.1.

The edges thus constructed are then partitioned into *natural graphs* according to the following principle: If an edge $\{x, y\}$ belongs to a natural graph, then so does some edge of form $\{\bar{x}, z\}$ and some edge of form $\{\bar{y}, w\}$. If a natural graph has an even number of edges, as on the left of Figure 5.1, it can be shown that in all optimal ancestral doubled genomes, if a gray edge, representing two adjacent vertices in the ancestor, has a vertex in this natural graph, then it necessarily connects to another vertex in the same natural graph. For natural graphs with an odd number of edges, which cannot be completed by adding pairs of edges, there are one or more ways of grouping them pairwise into *supernatural graphs*, as on the right of Figure 5.1. An optimal doubled ancestor exists such that if a gray edge has a vertex in this supernatural graph, then it connects to another vertex within the same supernatural graph. Thus the supernatural graphs may be completed one at a time.

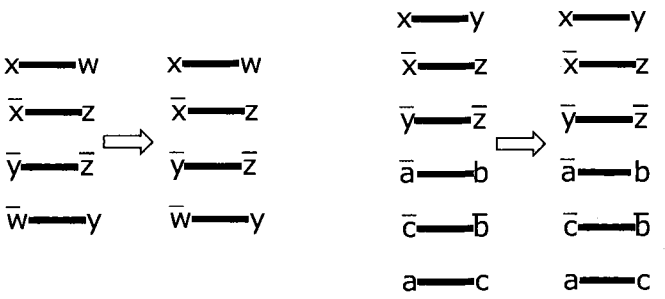

Figure 5.1: (left) Even-size natural graph completed by adding two pairs of gray edges. (right) Two odd-size natural graphs, containing x, y, z vertices and a, b, c vertices, respectively, combined into one supernatural graph so that three pairs of gray edges may be added.

An important detail in this construction is that before a gray edge is added during the completion of a supernatural graph, it must be checked to see that it would not inadvertently result in a circular chromosome. This involves inspection within this supernatural graph only. Key to the linear worst-case complexity of the halving algorithm is that this check may be made in constant time.

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.

5.3.3 Median.

Unlike the genomic distances and genome halving, which can all be calculated in linear time, the genome median problem, based either on d or d_B, is NP-hard [35, 46, 153]. The heuristics [26, 67] commonly used to analyze the problem search for reversals that will move a genome towards the other two. This is iterated as often as

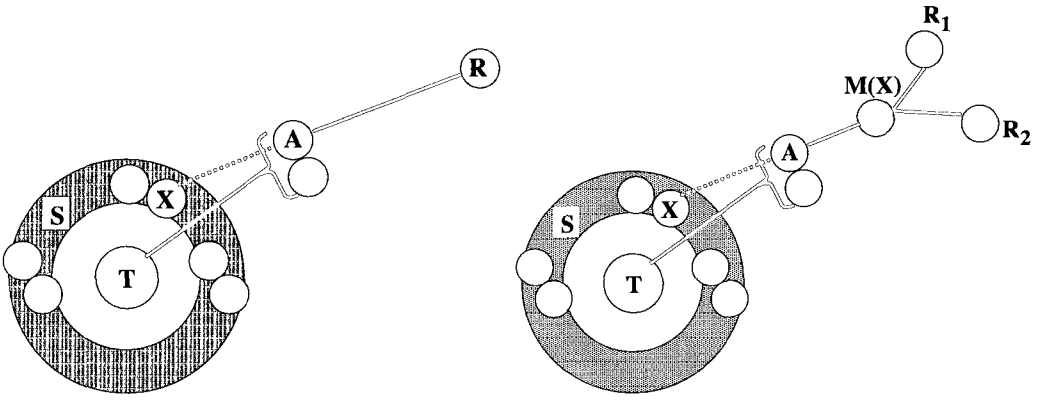

Figure 5.2: Halving a doubling descendent T, with one (R) or two (R_1, R_2) unduplicated outgroups. The double circles represent two copies of potential ancestral genomes, including solutions to the genome halving in **S**, and those on best trajectories between **S** and outgroups.

possible; otherwise one of the genomes is moved towards only one of the others without prejudicing its distance to the third, and the algorithm stops when all three genomes become identical. These algorithms become prohibitively costly with moderate n.

5.4 Previous work on GGH.

5.4.1 Guided genome halving with one outgroup.

Consider T and and a related unduplicated genome R with genes orthologous to $a_1, \cdots, a_n$. Our problem is to find an unduplicated genome A that minimizes, for some $A' \oplus A''$,

$$D(T, R) = d(R, A) + d(A' \oplus A'', T). \tag{5.4.1}$$

Our solution in [238], as on the left of Figure 5.2, is to generate the set **S** of genome halving solutions, then to focus on 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 an element $X \in \mathbf{S}'$ and the outgroups. First, each possible genome on one or more trajectories between X and R is examined in turn to see if it that decreases $D(T, R)$. If so, it is taken as the current best value of X. When no better X is found for any starting point in $\mathbf{S}'$ the current value is taken to be A.

In our experience, any more comprehensive search becomes computationally very costly, and very rarely finds a better solution.

When $\mathbf{S}'$ is so large that an exhaustive search for a local minimum becomes computationally too costly, or when it is too costly to generate all of $\mathbf{S}$ in order to find $\mathbf{S}'$, we may resort to sampling $\mathbf{S}$. In defining the gray edges in the supernatural graphs of Section 5.3.2, we generally have several choices at some of the steps. By randomizing this choice, we are effectively choosing a random sample of $X \in \mathbf{S}$.

5.4.2 Guided genome halving with two outgroups.

With reference to the right of Fig. 5.2, consider T and two unduplicated genomes R_1 and R_2 with genes orthologous to $a_1, \cdots, a_n$. Our problem here is to find a genome A and a median genome M for A, R_1 and R_2 that minimize

$$D(T, R_1, R_2) = d(R_1, M) + d(R_2, M) + d(A, M) + d(A' \oplus A'', T) \tag{5.4.2}$$

for some $A' \oplus A''$. Our solution in [174], as on the right of Fig. 5.2, was to generate the set $\mathbf{S}$ of solutions of the genome halving problem, then to focus on the subset $X \in \mathbf{S}' \subset \mathbf{S}$ where $d(R_1, M)+d(R_2, M)+d(X, M)$ is minimized, using our own implementation of the median heuristics mentioned in Section 5.3.3. Then we sought the A minimizing $D(T, R_1, R_2)$, heuristically, along all trajectories between all elements $X \in \mathbf{S}'$ and $M(X)$.

5.5 Complexity

We prove that GGH for one outgroup under the breakpoint distance d_B is NP-hard, using a reduction from the Breakpoint Median Problem. The latter is NP-hard, both for unichromosomal [46] and multichromosomal genomes (E. Tannier, personal communication).

We convert the breakpoint median problem on P, Q and R, three diploid genomes with the same genes, into an instance of GGH:

- Construct genome P_1 by appending superscript '1' to the symbol for each gene in genome P.
- Construct genome Q_2 by appending superscript '2' to the symbol for each gene in genome Q.
- Let $T = P_1 \oplus Q_2$. We will treat T as a doubling descendant. Superscripts '1' and '2' distinguish the two copies of a gene.
- Define an instance of GGH based on the doubling descendant T and the diploid outgroup R.

We prove that the solution of GGH for genomes T and R is also the solution of Breakpoint Median Problem on genomes P, Q and R:

Given any assignment of the '1' and '2' superscripts to the pairs of genes in T, a solution for GGH minimizes

$$B(T, R) = d_B(R, A) + d_B(A' \oplus A'', T). \tag{5.5.1}$$

where A' is a genome with one copy of each gene, labeled '1' or '2', and A'' is the

same as A' with all the '1' and '2' superscripts interchanged. A is the same genome without superscripts.

Lemma 1. *If we construct genome A_1 by appending superscript '1' to each gene in genome A, and A_2 by appending superscript '2' to each gene in genome A, then*

$$d_B(A_1 \oplus A_2, T) = d_B(A' \oplus A'', T). \tag{5.5.2}$$

Proof. Genomes A' and A'' form a solution to GGH. The sum $d_B(A' \oplus A'', T) + d_B(A, R)$ is minimized. Therefore

$$d_B(A' \oplus A'', T) + d_B(A, R) \leq d_B(A, R) + d_B(A_1 \oplus A_2, T). \tag{5.5.3}$$

Due to the construction of the genome T, each pair of adjacent elements in T must have the same superscript. This implies that for every adjacency that $A' \oplus A''$ has in common with genome T, the two adjacent terms must have same superscript too. Genome $A_1 \oplus A_2$ contains all these common adjacencies, which implies

$$d_B(A_1 \oplus A_2, T) \leq d_B(A' \oplus A'', T). \tag{5.5.4}$$

Thus $d_B(A_1 \oplus A_2, T) = d_B(A' \oplus A'', T)$. If A' and A'' form a solution of GGH, then A_1 and A_2 also constitute a solution with the same breakpoint distance. □

Lemma 2. *The breakpoint distance $d_B(A_1 \oplus A_2, T) = d_B(A, P) + d_B(A, Q)$.*

Proof. We constructed $T = P_1 \oplus Q_2$. The adjacencies in common between $A_1 \oplus A_2$ and T can be divided into two kinds:

- the common adjacencies between A_1 and P_1; and
- the common adjacencies between A_2 and Q_2.

Therefore $d_B(A_1 \oplus A_2, T) = d_B(A_1, P_1) + d_B(A_2, Q_2)$. Trivially, i.e., by simply ignoring superscripts, $d_B(A_1, P_1) = d_B(A, P)$ and $d_B(A_2, Q_2) = d_B(A, Q)$

□

Theorem 1. *Genome A, the solution of GGH for T and C, is also the solution of the Breakpoint Median Problem on genomes P,Q and R.*

Proof. From Lemma 2, $d_B(A_1 \oplus A_2, T) = d_B(A, P) + d_B(A, Q)$. Thus

$$d_B(A_1 \oplus A_2, T) + d_B(A, R) = d_B(A, P) + d_B(A, Q) + d_B(A, R). \tag{5.5.5}$$

There cannot be any other genome A^* such that $d_B(A^*, P) + d_B(A^*, Q) + d_B(A^*, R) < d_B(A, P) + d_B(A, Q) + d_B(A, R)$, because this A^* would then have the property that

$$d_B(A_1 \oplus A2, T) + d_B(A, R) > d_B(A_1^* \oplus A_2^*, T) + d_B(A^*, R), \tag{5.5.6}$$

a contradiction. Therefore genome G is the solution of the Breakpoint Median Problem on P,Q and R. □

Assuming the Breakpoint Median Problem for four genomes L, P, Q and R were also NP-hard, although we are not aware of any explicit proof, we could use the same method employed above to show that GGH with two outgroups is hard under the d_B distance.

We do not yet have corresponding proofs that GGH is NP-hard under the rearrangement distance d, but this is almost certainly the case since the breakpoint distance easier to compute than rearrangement distance, even though they are both $O(n)$. Note that the Reversals Median Problem for three or more (unichromosomal) genomes is NP-hard [35].

5.6 The new algorithms

The key idea in our improvement on the brute force algorithms is to combine information from both T and the outgroups in constructing the ancestor. It is important to take advantage of the common structure in T and the outgroups as early as possible, before it can be destroyed in the course of construction. To this end, we drop the practice of completing the all the gray edges in one supernatural graph before starting another. We simply look for elements of common structure and add gray edges accordingly, making sure at each step that no circular chromosomes are inadvertently created. It is still necessary to construct the supernatural graphs at the outset, both for the check against circular chromosomes and for technical reasons we omit here, having to do with chromosome ends.

Our approach requires only slight modifications from the context of a single outgroup to that of two outgroups. For that reason, we present a single algorithm for both, with the modifications for two outgroups in square brackets. Indeed, this presentation is suggestive of a generalization to three or more outgroups.

5.6.1 Paths.

By "path" we mean any connected succession of black and gray edges in a breakpoint graph, starting and terminating with a black edge. We represent each path by an unordered pair $(u, v) = (v, u)$ consisting of its current endpoints, though we keep track of all its vertices and edges. Initially, each black edge in T is a path, as is each black edge in R [or in each of R_1 and R_2].

5.6.2 Pathgroups.

A pathgroup Γ is an ordered triple [quadruple] of paths, two in T and one in R [one each in outgroups R_1 and R_2], where one endpoint of one of the paths in T is the duplicate of one endpoint of the other path in T and both are orthologous to one of the endpoints of the path in R [R_1 and R_2]. The other endpoints may be duplicates or orthologs to each other, or not. For the special case where the duplicates are end vertices, and the supernatural graph containing it has four end nodes, then the members of a pair of duplicate dummies must originate in different (odd length) natural graphs.

5.6.3 The algorithms

In adding pairs of gray edges to connect duplicate pairs of terms in the breakpoint graph of T versus $X' \oplus X''$, (which is being constructed), our approach is basically greedy, but with an important look-ahead. We can distinguish six priority levels among potential gray edges, i.e., potential adjacencies in the ancestor. Recall that in constructing the ancestor X to be close to the outgroups, such that $X' \oplus X''$ is simultaneously close to T, we must create as many cycles as possible in the breakpoint

graphs between X and the outgroups and in the breakpoint graph of $X' \oplus X''$ versus T.

1. Adding two gray edges would create two cycles in the breakpoint graph defined by T and $X' \oplus X''$, by closing two paths, as on the top of Figure 5.3. When this possibility exists, it must be realized, since it is an obligatory choice in any genome halving algorithm. It may or may not also create cycles in the breakpoint graph comparison of X with the outgroups.

2. Adding two gray edges would create three cycles, one for T and one for each of two outgroups.

3. Adding two gray edges would create two cycles, one for T and one for one outgroup, as in the middle of Figure 5.3.

4. Adding two gray edges would create one cycle for T but none for the outgroups. It would, however, create a higher priority pathgroup, e.g., Figure 5.3, bottom.

5. Adding two gray edges would create a cycle in the T versus $X' \oplus X''$ comparison, but none for the outgroups, nor would it create any higher priority pathgroup.

6. Each remaining path terminates in duplicate terms, which cannot be connected to form a cycle, since in $X' \oplus X''$ these must be on different (and identical) chromosomes. In supernatural graphs containing such paths, there is always another path and adding two gray edges between the endpoints of the two paths can create a cycle.

In not completing each supernatural graph before moving on to another, we lose the advantage in [51] of a constant time check against creating circular chromosomes. The worst case becomes a linear time check. In practice, this is a small liability, because the worst case scenario is seldom realized.

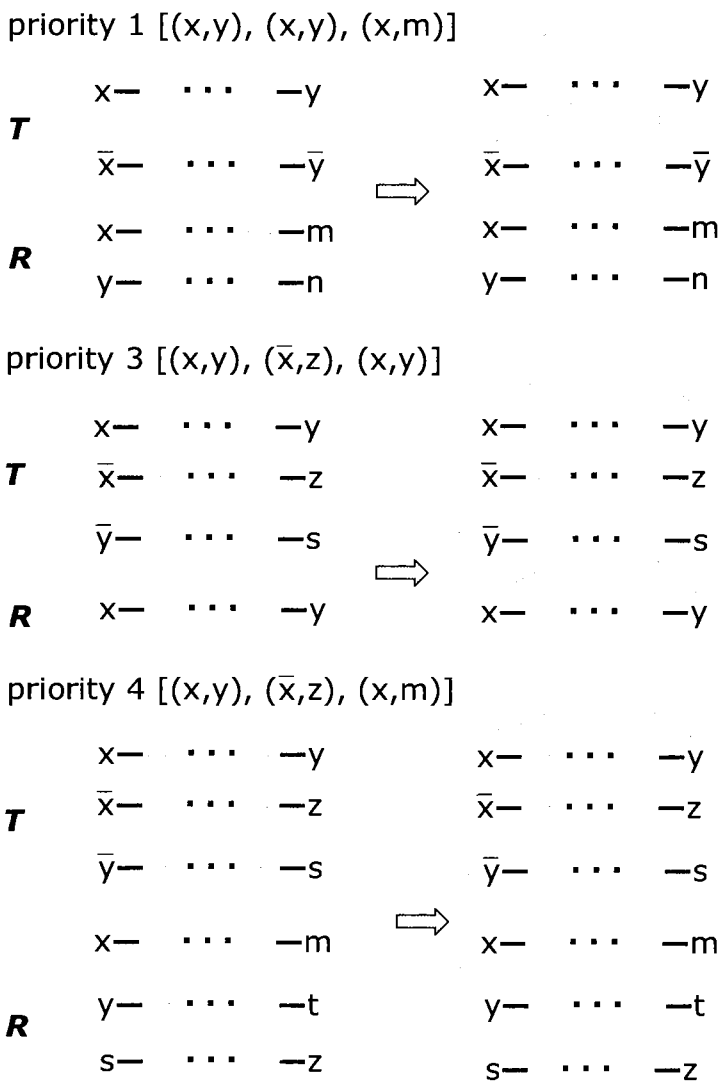

Figure 5.3: Priority levels of some pathgroups for GGH with one outgroup.

Algorithm GGH:

Guided Genome Halving with One [Two] Outgroups

Input. Two [three] genomes:

duplication descendant T, outgroup R $[R_1, R_2]$.

Output. Genome X, a halving solution of T, minimizing

$d(X' \oplus X'', T) + d(X, R)$

$[d(X' \oplus X'', T) + d(X, R_1) + d(X, R_2)]$.

Initialize paths (black edges) in T and R [in R_1 and R_2].

Construct supernatural graphs.

Construct two pathgroups for each gene g in R [in R_1],

one based on g_t, the other on g_h.

If number of chromosomes in T is odd,

add pathgroup with two paths of form (end, end).

While there remains at least one pathgroup

For each pathgroup:

$((x, y), (\bar{x}, z), (x, m))$, or $[((x, y), (\bar{x}, z), (x, m), (x, n))]$

classify it by case and priority,

and find a pathgroup Γ that has the highest priority.

Case 1: $\bar{x} \neq y$, and adding xy and $\bar{x}\bar{y}$ would not create a circular chromosome.

Priority 1: $z = \bar{y}$.

[Priority 2: $y = m = n$.]

Priority 3: $y = m$ [or $y = n$].

Priority 4: adding xy and $\bar{x}\bar{y}$ would create a pathgroup with priority 2 or 3.

Priority 5: None of 1, [2], 3 or 4.

Case 2: $\bar{x} \neq y$, and adding $x\bar{z}$ and $\bar{x}z$ would not create a circular chromosome.

[Priority 2: $z = m = n$.]

Priority 3: $z = m$ [or $z = n$].

Priority 4: adding $x\bar{z}$ and $\bar{x}z$ would create a pathgroup with priority 2 or 3.

Priority 5: None of [2], 3 or 4.

Case 3:$\bar{x} = y$.

Priority 6:

If Γ is Case 1, **addGrayEdge**$(xy, \bar{x}\bar{y})$.

If Γ is Case 2, **addGrayEdge**$(x\bar{z}, \bar{x}z)$.

If Γ is Case 3, find some

$W = ((w, \bar{w}), (\bar{w}, w), (w, s))$ or $[((w, \bar{w}), (\bar{w}, w), (w, s)(w, t))]$

in the same supernatural graph and **addGrayEdge**$(xw, \bar{x}\bar{w})$.

Algorithm: addGrayEdge$(rt, \bar{r}\bar{t})$

Add gray edges $rt, \bar{r}\bar{t}$ to partially completed genome $X" \oplus X''$.

Add gray edge rt to partially completed genome X.

Update paths in pathgroups that are affected by the new gray edges.

Remove pathgroups that start with r and t.

Once the GGH algorithm is terminated, we undertake the local search described in Sections 5.4.1 and 5.4.2 to see if we can improve X by allowing it to move out of **S** on a trajectory towards R.

Table 5.1: Performance comparison of sampling method and guided halving algorithm in the case of one outgroup. Sample size 2000 for the sampling method. $R-T$ represents the outgroup and doubling descendant. n is the number of genes available in that pair of genomes, with two copies in T. $d_{t,x\oplus x} = d(T,X'\oplus X'')$ is the doubling distance, constant over all analyses. $\bar{d}_{x,r} = \bar{d}(X,R)$ represents the average, over all samples, of the distance estimate between the ancestor, just before doubling, and the outgroup, and the adjacent entry $d_{\min} = \min_{\text{sample}} d(X,R)$ is the minimum found. Δ_A is the improvement over $d(T,X'\oplus X'') + d(X,R)$ due to local searching, allowing A to be found outside the set of halving solutions. $d_{a,a^*} = d(A,A^*)$ is the distance between the inferred ancestor and the "ground truth". Time is measured in minutes, for 2000 samples of unrestricted halving or for one GGH run.

halving analysis			sampling method						guided halving				
$R-T$	$2n$	$d_{t,x\oplus x}$	$\bar{d}_{x,r}$	$d_{\min}$	$d_{a,r}$	Δ_A	d_{a,a^*}	time	$d_{x,r}$	$d_{a,r}$	Δ_A	d_{a,a^*}	time
AG-CG	538	186	204	196	180	-16	156	37	153	153	0	120	2.3
AG-SC	1012	119	237	229	208	-21	53	158	184	183	-1	32	5.3
KL-CG	546	186	210	203	184	-19	154	50	160	160	0	120	3.5
KL-SC	1026	122	241	232	216	-16	51	140	197	197	0	39	6.1
KW-CG	542	188	247	238	230	-8	167	26	216	215	-1	142	3.3
KW-SC	994	121	364	355	350	-5	70	72	325	323	-2	41	5.1
A*-CG	600	199	183	169	129	-40	129	81	84	84	0	84	1.5
A*-SC	1062	124	79	70	37	-33	37	114	5	5	0	5	0.3
AG-V	576	61	157	151	149	-2	54	12	148	148	0	51	0.9
KL-V	584	62	167	160	158	-2	53	12	157	157	0	51	0.9
KW-V	582	62	224	218	215	-3	52	13	212	212	0	51	1.0
A*-V	600	62	57	49	39	-10	39	14	29	29	0	29	0.2

5.7 Genome doubling in yeast

Wolfe and Shields [227] discovered an ancient genome doubling in the ancestry of *Saccharomyces cerevisiae* in 1997 after this organism became the first eukaryote to have its genome sequenced [62]. According to [95], the recently sequenced *Candida glabrata* [46] shares this doubled ancestor. We extracted data from YGOB (Yeast Genome Browser) [31], on the orders and orientation of the 600 genes (300 pairs) identified as duplicates in both genomes.

The Yeast Gene Order Browser (YGOB) [31] contains complete gene orders and orthology identification among the five yeast species depicted in Figure 5.4: the two

Table 5.2: Results of guided halving algorithm in the case of two outgroups. Median cost refers to the sum of the three distances, from R_1, R_2 and the inferred ancestor X or A, to the median. The objective is $d(T, X' \oplus X'')$ + median cost. Δ_A is the improvement of $d(T,A \oplus A)$ + median cost over $d(T, X' \oplus X'')$ + median cost due to local searching, allowing A to move outside the set of halving solutions. Time in minutes.

R_1-R_2-T	n	$d(T,X' \oplus X'')$	median cost	$d(T,A' \oplus A'')$	median cost	Δ_A	$d(A, A^*)$	time
AG-KL-SC	497	117	364	117	361	-3	40	131
AG-KW-SC	478	116	502	116	498	-4	41	204
KL-KW-SC	471	121	518	121	516	-2	48	217
AG-KL-CG	265	183	300	183	297	-3	124	48
AG-KW-CG	261	184	362	184	361	-1	138	55
KL-KW-CG	259	184	368	184	366	-2	136	62
AG-KL-V	283	61	278	61	275	-3	47	38
AG-KW-V	280	61	340	62	339	0	51	41
KL-KW-V	277	62	354	62	352	-2	54	54

descendents of the above-mentioned ancient genome duplication event, *Saccharomyces cerevisiae* and *Candida glabrata*, and three species that diverged before this event, *Ashbya gossypii, Kluyveromyces waltii* and *Kluyveromyces lactis.* For the ancient tetraploids, YGOB includes a reconstruction of the ancestral genome. We abbreviate these six genomes as SC, CG, AG, KW, KL and A*, respectively. In addition, we construct an ancestral doubled descendant V lying on a shortest rearrangement trajectory from SC to CG, satisfying the criterion that its halving distance is minimal [240]. We take the ancestor A* as "ground truth" and see how close we can approach it using the sampling method and the guided halving method, with various combinations of doubling descendants and unduplicated genomes.

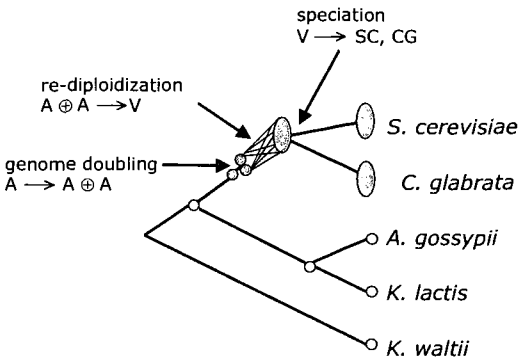

Figure 5.4: Phylogeny of yeasts in YGOB. Whole genome doubling event giving rise to ancestor of *S. cerevisiae* and *C. glabrata* indicated, followed by rediploidization and speciation and the divergence of these two species.

5.8 Results

Table 5.1 compares the results, before and after local optimization, of the guided halving algorithm and the sampling approach on 12 pairs of genomes, the three doubling descendants SC, CG and V, each versus the four unduplicated genomes AG, KL, KW and A*. Recall that V and A* are themselves analytical constructs, the former representing the most recent common ancestor of SC and CG, and the latter the ancestral genome at the moment of doubling.

The first observations are methodological. In all 12 cases guided halving results in an X closer to R than in any of 2000 samples of unrestricted halving. If computing time were no obstacle, the sampling method would be exhaustive and exact, and hence always at least as good as guided halving. The fact that none of the 12 analyses produced a "lucky" sample as good as or better than GGH, suggests that we would need a sample size of 25,000 at the very least, and perhaps one or more orders of magnitude larger, to bring the accuracy of sampling method to the level of guided halving, but this would require thousands of hours or more for our entire data set versus less than 30 minutes with guided halving.

The fact that the results of the sampling method are improved by local searching, usually substantially, in all 12 cases, whereas guided halving produces genomes already at or very close to a minimum (albeit possibly local) of the objective function, is another measure of the superior performance of the latter.

Note that aside from the three cases where the ground truth ancestor A^* plays the role of outgroup, this genome is not directly involved in the analysis. It is of great interest, then, from the biological viewpoint, that in all cases, guided halving produces an ancestor A closer to A^* than the sampling method. Moreover, when using A^* as an outgroup for the halving of SC, the analysis reconstructs something very close to A^*, i.e., where $d(A, A^*)$ is only 5. This attests to the internal coherence of the method: the SC evidence was predominant in the original construction of A^* [31].

Turning to the case of two outgroups, we first point out that the sampling approach becomes infeasible when even a moderate number of analyses are undertaken. This is due to the relatively lengthy time (sometimes more than two hours) required to compute the median cost, i.e., the sum of the three distances, from R_1, R_2 and the inferred ancestor X, to the median. (The halving algorithm alone, and even guided halving, never takes more than 2 or 3 minutes.) This is not an obstacle to the guided halving method because the median need to be calculated just once, instead of the thousands of times for the sampling approach. Table 5.2 shows the result of halving guided by two outgroups, using all combinations of two of AG, KL and KW versus each of SC, CG and V.

In general, we note no advantage of using two outgroups over one, in that $d(A, A^*)$ with two outgroups is not as good as $d(A, A^*)$ for the better of the two used alone. The exception is the comparison of KL and AG with V. Thus it seems, at least with these data, that the more remote outgroup contributes little more than noise to the reconstruction guided by the closer outgroup. This result may be due to the great

discrepancy in the phylogenetic divergence between the doubled genomes and KW compared to the divergence between the former and AG or KL, and may not carry over to other data sets.

Two observations: first, the improvement due to local search is relatively small, though larger than guided halving with one outgroup. Second, though our analyses did find some A outside of $\mathbf{S}$ that minimized $D(T, R_1, R_2)$, in each such case there was also a solution (the one entered in Table 5.2) with $A \in \mathbf{S}$.

To investigate to what extent differences between the doubling descendants and among the outgroups are reflected in the reconstructed ancestor genome A, we undertook Gower's principal coordinates analysis [64] of the 21 versions of A described in Tables 5.1 and 5.2, as well as A^* itself. We used the implementation of this analysis available as *cmdscale* in the R environment [161], applied to the 22 genomic distance matrix.

Figure 5.5 depicts the results of a three-dimensional principal coordinates analysis. We note first that the first two dimensions basically distinguish among the doubling descendants, first classifying SC and V together versus CG, and then distinguishing between SC and V. The third dimension distinguishes between the genomes in which KW was the outgroup and those in which only AG and/or KL were outgroups. As we would expect, all the genomes with A^* as the outgroup or as one of two outgroups, are closer to the "true" ancestor A^* than when some other outgroup is used instead. Nevertheless, other outgroups, such as AG, also help guide the reconstruction to fairly close approximations of A^*. On the other hand, constructions guided by CG are all very far from A^*, and those involving KW tend to be somewhat farther than those guided by AG and KL. The latter observation is consistent with the known highly rearranged nature of CG, and with the relatively distant evolutionary relationship between KW and A^*, as can be seen in Figure 5.4.

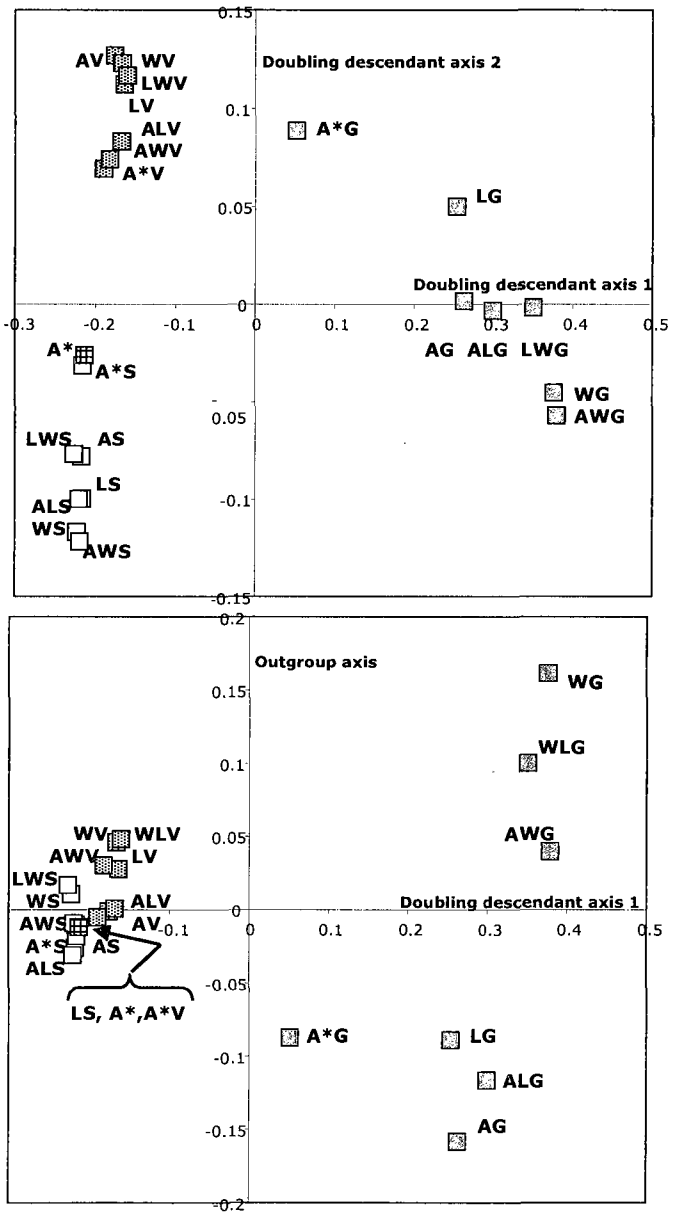

Figure 5.5: First three dimensions of principal coordinate analysis of distances among 22 inferences of ancestral genome, based on different configurations of outgroups. Up: dimensions 1 and 2. Down: dimensions 1 and 3. Dimension labels assigned subjectively *after* the analysis. Genomes SC, CG, AG, KL and KW further abbreviated in displays to S, G, A (not to be confused with A for ancestor elsewhere in the text, nor with A^*), L and W, respectively.

5.9 Discussion

We have focused on the two main concerns of genome halving, the multiplicity and the diversity of solutions, and the difficulty of assessing the accuracy of the results with real data. Guided genome halving was previously shown to drastically reduce the non-uniqueness inherent in unrestricted halving. This is carried further by GGH, which achieves much greater accuracy with much less computational effort.

An important indication of the precision of the reconstruction is its ability with some of the data to come very close to the manually reconstructed ancestor A.

Nevertheless, these results remind us of the uncertainties inherent in historical reconstruction. Some of this is possibly due to the "noise" of mistaken paralogy identification, especially in highly rearranged genomes such as *Candida glabrata*. Future work will attempt to attenuate this noise using the techniques of [239] and [39].

The significance of halving results depends on what proportion of the doubling descendant T can be identified as duplicated genes. Our analysis does not attempt to situate the ancestors of genes present in only one copy in T, and these will often form the majority. Ongoing work exploits the syntenic relationships between these genes, the duplicated ones, and their orthologs in the outgroups.

Acknowledgements

We thank Ken Wolfe, Kevin Byrne and Jonathan Gordon for encouragement and for valuable information. We also thank Howard Bussey, Eric Tannier and Robert Warren for helpful discussions. Research supported in part by a grant to DS from the Natural Sciences and Engineering Research Council of Canada (NSERC). DS holds the Canada Research Chair in Mathematical Genomics.

Chapter 6

Polyploidy and angiosperm diversification

D. Soltis, J.Leebens-Mack, C.D. Bella, A. Paterson, V.A. Albert, C. Zheng, D. Sankoff and P.S. Soltis, 2008. American Journal of Botany 96:336–348.

Dr. Sankoff and I participated in discussions with Dr. Soltis, Leebens-Mack and Albert during the writing of this review, and we contributed the section "Reconstructing ancestral genomes prior to genome duplication" as well as Fig. 6.3, which I prepared using the algorithm from Chapter 5 of this thesis.

Abstract

Polyploidy has long been recognized as a major force in angiosperm evolution. Recent genomic investigations not only indicate that polyploidy is ubiquitous among angiosperms, but also suggest several ancient genome-doubling events. These include ancient whole genome duplication (WGD) events in basal angiosperm lineages, as well as a proposed paleohexaploid event that may have occurred close to the eudicot divergence. However, there is currently no evidence for WGD in *Amborella*, the putative sister species to other extant angiosperms. The question is no longer "What proportion of angiosperms are polyploid?", but "How many episodes of polyploidy characterize any given lineage?" New algorithms provide promise that ancestral genomes can be reconstructed for deep divergences (e.g., it may be possible to reconstruct the ancestral eudicot or even the ancestral angiosperm genome). Comparisons of diversification rates suggest that genome doubling may have led to a dramatic increase in species richness in several angiosperm lineages, including Poaceae, Solanaceae, Fabaceae, and Brassicaceae. However, additional genomic studies are needed to pinpoint the exact phylogenetic placement of the ancient polyploidy events within these lineages and to determine when novel genes resulting from polyploidy have enabled adaptive radiations.

6.1 Introduction

As the papers in this issue and other recent reviews illustrate [45, 53, 54, 83, 88, 207], through collective progress in paleobotany, phylogeny, and studies of the evolution of plant development, enormous progress has been made in understanding Darwin's "abominable mystery" – the "rapid rise and early diversification" of the angiosperms. For example, recent phylogenetic analyses suggest that following initial, early "experiments", angiosperms radiated rapidly ($\leq$ 5 million years), yielding the five extant lineages of Mesangiospermae (sensu [34], magnoliids and Chloranthaceae as sisters to a clade of monocots and eudicots + Ceratophyllaceae) [138, 166]. This rapid radiation ultimately produced ca. 97% of all angiosperm species. Here we emphasize that polyploidy, often referred to in the genomics literature as whole-genome duplication (WGD), has played a dramatic role in the diversification of most, if not all, eukaryotic lineages, perhaps most impressively within the angiosperms.

Researchers have long recognized that polyploidy is an inseparable part of angiosperm biology. Polyploidy in angiosperms has been studied for a century, dating to the work of De Vries [58, 122] and to early interest in a putative chromosome duplication in maize (*Zea mays*) [96]. Early reviews of polyploidy in plants included [37, 42, 119, 142, 200–203]. Following the work of Stebbins [199–201] in particular, polyploidy became a major focus of biosystematic research. As a result, plant scientists have long recognized that polyploid lineages may have complex relationships with each other and their diploid ancestors, making application of species concepts problematic [165, 191].

Fueled in part by evidence for ancient polyploidy in genome sequences of *Arabidopsis* [23, 24, 27, 150, 183, 218], and more recently published angiosperm genome sequences (discussed later), the past decade has seen a dramatic resurgence in the

study of polyploidy [3,104,189,209,222]. There has been renewed interest in the mechanisms of polyploid formation and establishment [75, 162, 163], the frequency of recurrent polyploidization [187,196], the ecological effects of plant polyploidy [211,212], and the genetic, epigenetic, chromosomal, and genomic consequences of polyploidization [3, 27, 105, 106, 118, 146, 152, 164, 206]. This research has resulted in major modifications to many of the traditional tenets of polyploid evolution.

Despite numerous attempts over the past 70 years to estimate the frequency of polyploidy in plants, it has proved difficult to determine the actual frequency of the process in various plant lineages. Angiosperms, in particular, have received much attention. Using chromosome numbers and hypotheses for the presumed dividing line between "diploid" and "polyploid" chromosome numbers, many estimates of the frequency of polyploidy in flowering plants have been made. However, these estimates have varied depending on the base chromosome number used, as well as on the sample of taxa considered. Müntzing [142] and Darlington [42] speculated that about one half of all angiosperm species were polyploid, while Stebbins [201] estimated the frequency of polyploidy to be 30 – 35% in angiosperms. Grant [65, 66], basing his estimate of the frequency of polyploidy in angiosperms on chromosome numbers for 17138 species available in 1955, hypothesized that flowering plants with haploid chromosome numbers of $n = 14$ or higher were of polyploid origin. Using this cut-off point, Grant inferred that 47% of all flowering plants were of polyploid origin and proposed that 58% of monocots and 43% of "dicots" were polyploid. Goldblatt [63] suggested that Grant's [65] estimate was too conservative; he thought that taxa with chromosome numbers above $n = 9$ or 10 would have had polyploidy in their evolutionary history. Using these lower numbers, he calculated that at least 70%, and perhaps 80%, of monocots are of polyploid origin. Lewis [112] applied a similar approach to "dicots" and estimated that 70 – 80% were polyploid. However, one problem with these sorts

of approaches is that they do not accommodate reductions in chromosome number that occurred rapidly on an evolutionary scale. For example, species of *Sorghum* with $n = 5$ are recently derived from taxa with $n = 10$ [198], and *Arabidopsis thaliana* with $n = 5$ appears to have resulted from three chromosomal condensations since its divergence from its common ancestor with *A. lyrata* and the more distantly related *Capsella rubella* [89, 94, 232].

Masterson [130] compared leaf guard cell size in fossil and extant taxa from a few angiosperm families (Platanaceae, Lauraceae, Magnoliaceae) to estimate polyploid occurrence through time. Because guard cell size is often much larger in polyploids than in diploids, this provided a gross estimate of whether the fossil taxa were diploid (smaller guard cells than extant taxa) or polyploid (the same or larger guard cell sizes vs. extant species). From these comparisons, Masterson (1994) estimated that 70% of all angiosperms had experienced one or more episodes of polyploidy in their ancestry. Using an innovative approach for estimating the incidence of polyploidy based on the distribution of haploid chromosome numbers, Otto and Whitton [147] suggested that roughly 2–4% of speciation events in angiosperms may have involved polyploidy. As a result, they suggested that "polyploidization may be the single most common mechanism of sympatric speciation in plants".

The older estimates of the frequency of polyploidy discussed tend to confound the actual process of polyploidization with the result (i.e., polyploid lineages). Actual WGD events should be far less frequent than particular chromosome numbers suggestive of polyploidy in any given angiosperm species. As reviewed next, through the interplay of genomic and phylogenetic approaches, we are on the verge of determining the frequency of ancient polyploidy events throughout angiosperm history.

6.2 A genomics approach to polyploidy

Summary of genomes sequenced to date –Investigations of completely sequenced nuclear genomes have dramatically altered the polyploidy paradigm. They have revealed that flowering plants, and perhaps all eukaryotes, possess genomes with considerable gene redundancy, much of which is the result of (ancient) WGDs. Those entire angiosperm genomes that have been completely sequenced to date – *Oryza sativa* [151] (rice, Poaceae), *Arabidopsis thaliana* [23, 27, 150, 183, 218] (Brassicaceae), *Populus trichocarpa* (poplar, Salicaceae) [213], *Vitis vinifera* (grape, Vitaceae) [81, 216], and most recently *Carica papaya* (papaya, Caricaceae) [135] – all show evidence of WGD events.

Complete sequencing of the very small genome (for angiosperms, 157 Mb [16]) of *Arabidopsis thaliana* (Brassicacae [10, 77]) revealed numerous duplicate genes and suggested two or three rounds of genome-wide duplication [10,23,27,150,183,215,218], corroborating early suspicions based on genetic mapping [90, 133]. Analyses of the complete genome sequence of *A. thaliana* suggested three ancient polyploidy events [24, 27]. Bowers et al. [27] proposed that one of these events (termed α) may have occurred within Brassicales and suggested the other two (termed β and γ) could be considerably older (but this view has now been revised; see Fig 6.1 [81, 167]; and discussed later).

Sequencing of the nuclear genome of rice (*Oryza sativa*; Poaceae) again suggested ancient polyploidy in a plant with a chromosome number ($n = 12$) considered by many to be a model "diploid" [151, 233], although secondary associations (loose pairing at meiosis) among homologous and nonhomologous chromosomes have long been known [100]. The exact phylogenetic placement of rice's WGD event is unclear; Paterson et al. [151] estimated that the WGD occurred after the divergence of Poales from other

monocot orders but before the divergence of the major cereals from one another (see later). However, there is also equivocal evidence of a more ancient genome-doubling event in rice that might characterize all monocots [216]. The possibility of an ancient genome doubling in monocots is gaining support from analysis of recently sequenced genomes and new analytical methods (H. Tang, J. Bowers, X. Wang, University of Georgia; A. Paterson, unpublished data).

Sequencing of the *Populus* (Salicaceae) genome also revealed an independent WGD event after the divergence of the eurosid I and eurosid II lineages (Fabidae and Malvidae [34]), but before the divergence of *Salix* and *Populus* [213]. Ancient polyploidy in Salicaceae (both *Populus* and *Salix*) was suggested by Stebbins [201] and has been supported by isozyme data [187]. It remains to be determined if this genome doubling event will be evident in other lineages of the now expanded Salicaceae [9]. A second, older duplication is shared by the *Populus* and *Arabidopsis* lineages [213].

Based on analyses of the complete genome sequence of *Vitis*, Jaillon et al. [81] suggested that the common ancestor of *Vitis*, *Populus*, and *Arabidopsis* was an ancient hexaploid (this is now considered the γ event) that possibly arose after the divergence of the monocot and eudicot lineages [167]. The exact placement of the γ event remains unclear (hence the ambiguity in its placement in Fig. 6.1); γ may have occurred before the monocot – eudicot divergence. However, a paleohexaploid signature is not apparent in rice, favoring a placement after the divergence of monocots and eudicots. Following the formation of this hypothetical paleohexaploid, there were subsequent, distinct, genome-wide duplication events in the Brassicales and *Populus* lineages (Fig. 6.1). Velasco et al. [216] proposed, in contrast to Jaillon et al. [81] , that the second WGD event evident in *Vitis* was more recent (Fig. 6.1). Velasco et al. proposed that three genome-wide duplications occurred in the ancestors of both *Arabidopsis* and *Populus*: one shared by all eudicots (and perhaps also monocots),

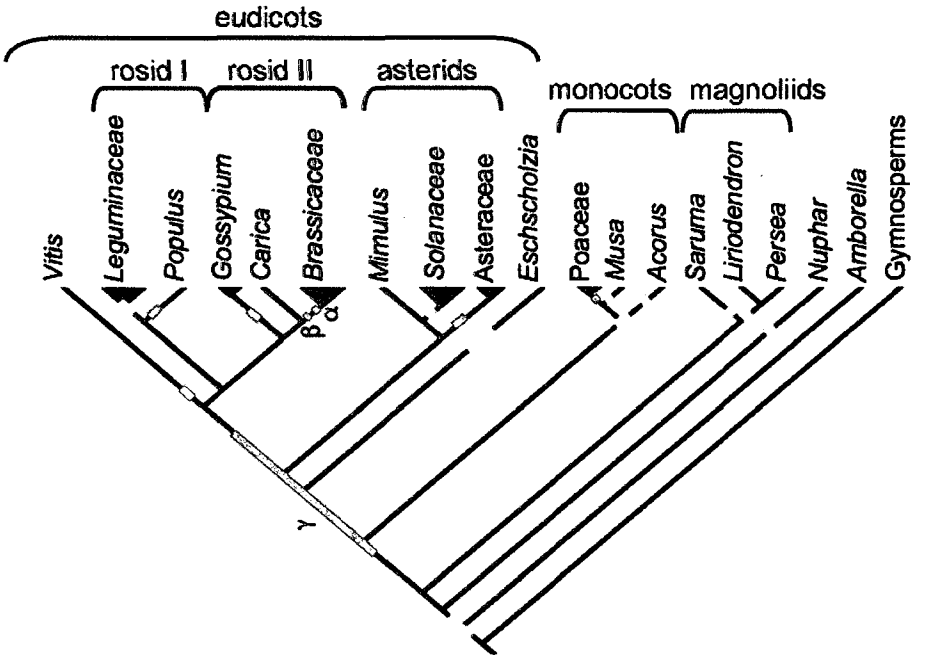

Figure 6.1: Simplified summary tree for angiosperms (following the general topology of D. Soltis et al. [188], with modifications reflecting more recent analyses, including Jansen et al. [83] ; Moore et al. [138]), depicting putative locations of genome duplication events now inferred for flowering plants relative to major lineages or species with sequenced nuclear genomes or substantial expressed sequence tag data. Tree redrawn to show *Vitis* sister to all other rosids [190], and *Carica* is placed as sister to Brassicaceae; *Carica* and Brassicaceae represent Brassicales. The α, β, and γ duplication events of Bowers et al. [27] have been added to this topology to match what appears to be the best location based on other analyses [206]. The timing and nature of the γ event are controversial with estimates of a tetraploid [216] or hexaploid [81] progenitor ranging from before the divergence of the monocot and eudicot lineages to the last common ancestor of all extant rosids. The α event likely occurred within Brassicaceae. The position of β is less certain; it may also have occurred within Brassicaceae, or perhaps earlier in Brassicales (see Fig 6.5). Blue bars represent genome duplications inferred from comparative analyses of physical and/or genetic maps. The red bar on the *Vitis* lineage represents the Velasco et al. [216] (2007) inference of a second distinct tetraploidy evident in the *Vitis* genome (but see Jaillon et al., [81] for a different interpretation). Yellow bars represent genome-wide duplication events inferred from analyses of paralog pairs found in large EST sets [25, 41].

a second WGD shared by *Arabidopsis* and *Populus* (cf. Fig. 6.1), and additional, single WGDs unique to each lineage. Velasco et al. [216] proposed that *Vitis* has the ancient genome wide duplication shared by all eudicots, as well as a lineage-specific duplication that may be the result of hybridization. The most ancient of these events is now thought to correspond to the γ event of Bowers et al. [27].

The nuclear genome of the eudicot *Carica papaya* (Caricaceae, Brassicales) has recently been sequenced [135]. *Carica* arose from a deep split in the order and is distantly related to Brassicaceae, the family that contains both *Arabidopsis* and *Brassica* (discussed later). There are no recent genome duplications evident in *Carica* and nothing corresponding to the *Arabidopsis* α or β events proposed by Bowers et al. [27]. Because Caricaceae are estimated to have diverged from the Brassicaceae lineage ca. 72 million years ago [135], the absence of recent genome duplication is in conflict with prior estimates of a much earlier age for the β genome duplication, which Bowers et al. [27] suggested could correspond to the origin of the eudicots. This apparent inconsistency is under investigation (J. Bowers, University of Georgia; A. Paterson, unpublished data). However, *Carica* shows evidence of γ, the early event detected in *Vitis, Populus*, and *Arabidopsis* that is now interpreted by Jaillon et al. [81] as a possible ancient triplication event. *Carica* has a reduced gene number, about 10% fewer genes than *Arabidopsis*, which may be largely accounted for by a paucity of genome duplications relative to other sequenced angiosperms. The data now suggest that both the α and β events occurred within Brassicales and are younger than originally proposed by Bowers et al. [27]. The α event likely occurred somewhere within Brassicaceae. However, the position of β is less clear; it may have occurred within Brassicales at some point after the divergence of Caricaceae (Fig. 6.1; more detail on these duplications is given later).

Although it is now clear that genome-wide duplications have occurred frequently

in angiosperms (Fig. 6.1), the exact number and precise phylogenetic placement of most of these events remain uncertain. Clarifying the timing and placement of these duplication events, particularly the older ones, will require additional genome sequencing, particularly of one or more basal angiosperms, providing the opportunity for ancestral-state reconstruction of numerous genomic characters [79,159]. Nevertheless, even as the number of angiosperm genome sequences increases, the oldest genome duplications will remain difficult to detect because recurrent polyploidy events, together with gene loss, chromosomal inversions, and translocations following genome duplication may obscure evidence of the earliest events in angiosperm history.

Because phylogenetic analyses [79,102,138,190,195] identify *Amborella trichopoda* (Amborellaceae) as the probable single sister species to all other living flowering plants, complete sequencing of the *Amborella* nuclear genome offers the opportunity to clarify how gene families and genome structure may have evolved within angiosperms [103, 192]. In addition, because the branching point for *Amborella* is located "between" gymnosperms and all other angiosperms, a genome sequence for *Amborella* could potentially help characterize genomic features that distinguish angiosperms from other seed plants. An *Amborella* nuclear genome sequence would facilitate efforts to reconstruct characteristics of the "ancestral angiosperm" [103,192]–the most recent common ancestor of the crown clade.

EST data and lineage-specific duplications – In addition to complete sequencing of the nuclear genome, other major sources of genomic data are the numerous expressed sequence tags (ESTs) that are now available for many flowering plants. A recent survey of GenBank revealed over 50 million ESTs, representing diverse eukaryotic lineages. The thousands of ESTs available for many plants provide a useful "snapshot" of each genome. The rapidly growing EST data sets for diverse angiosperms can be employed to assess ancient polyploidy using a genomics method to estimate

whether a genome duplication event may have occurred in a lineage, as well as the approximate age of these gene duplication events [124]. This method evaluates the frequency distribution of persite synonymous divergence levels (K_s) for pairs of duplicate genes. A genome-wide duplication event results in thousands of paralogous pairs – all simultaneously duplicated. Evidence of past genome duplications can be seen as peaks in the distribution of K_s values for sampled paralogous pairs [25, 41, 124, 178]. Importantly, this method does not require information on the position of genes within the genome, and therefore it can be applied to any species for which there are moderate to large EST sets. However, caution must be used in interpreting K_s values, and it is important to use statistical testing to identify significant deviations from background duplication process [41, 178]. There are clear examples in which well-known genome duplication events were not detected in K_s distributions [25, 151]. As duplicated genes undergo divergence and death processes, it becomes increasingly likely that polyploidy events will go undetected [41]. Finally, processes such as a sudden proliferation of tandem gene duplications [25] or a single chromosome duplication could produce a K_s peak through a large-scale process distinct from a true polyploid event.

We summarize the many genomes thought to be ancient polyploids using analyses of complete genome sequences, as well as ESTs, on a modified summary tree for angiosperms (Fig. 6.1). Genetic investigation of other taxa using other methods suggests additional ancient polyploidy events. For example, "diploid" members of Brassica are, at the least, ancient tetraploids [90, 98, 160] (see Fig. 6.1) and perhaps ancient hexaploids based on analyses of linkage maps with a number of genes clearly represented multiple times [97, 121]. There are also other lineage-specific duplications (see later).

Blanc and Wolfe [25] investigated 14 model plant species (mostly crop species

with known recent polyploid history) for which large EST data sets were available and found spikes in the distributions of older paralogous pairs (with higher K_s values) in nine species, including *Zea* (maize), *Glycine* (soybean), *Gossypium* (cotton), and *Solanum* (tomato and potato). Schlueter et al. [178] similarly employed the analysis of K_s distributions to sets of paralogous pairs identified in large EST data sets for eight major crop species, including *Glycine*, *Medicago* (alfalfa), *Solanum, Zea, Sorghum, Oryza*, and *Hordeum* (barley), and inferred multiple independent genome duplications in Fabaceae, Solanaceae, and Poaceae over the last 14 – 60 million years.

When the K_s approach was applied to ESTs from a suite of basal angiosperms via the Floral Genome Project [5], evidence was found for episodes of ancient genomewide duplication in *Nuphar advena* (Nymphaeaceae), *Persea americana* (avocado; Lauraceae), *Liriodendron tulipifera* (yellow poplar or tulip tree; Magnoliaceae), and *Saruma henryi* (Aristolochiaceae) [41]. In addition, Cui et al. [41] detected independent genome duplications in the basal eudicot *Eschscholzia californica* (California poppy; Papaveraceae) and the basal monocot *Acorus americanus* (sweet flag: Acoraceae), both of which were distinct from duplications documented for core eudicots and Poaceae (Fig. 6.1).

More than one genome-wide duplication event is evident in *Nuphar* (Nymphaeaceae). One of these is likely restricted to Nymphaeaceae (Nymphaeales), but another may correspond to the oldest duplication so far discovered in angiosperms – the latter may date to the common ancestor of all angiosperms except *Amborella*, which so far lacks evidence of ancient polyploidy [41] (see later) (Fig. 6.1). Analysis of K_s values also provided evidence for genome-wide duplication in both *Persea* (Lauraceae) and *Liriodendron* (Magnoliaceae). One of these is shared by both families, corroborating evidence based on isozyme data [186] for an ancient polyploidy event (about 100 Mya) in their common ancestor. There is weak although inconclusive evidence

of a still older WGD in *Persea*, which may correspond to the event suggested for the common ancestor of all angiosperms except *Amborella*. Alternatively, this WGD could perhaps even predate the angiosperms (see later). Testing these and similar hypotheses will minimally require comprehensive transcriptome sequencing for additional basal angiosperms and a complete *Amborella* genome sequence [192].

Are all angiosperms of ancient polyploid origin? – All angiosperm nuclear genomes studied except that of *Amborella* have yielded evidence for multiple genome-wide duplications (Fig. 6.1). In contrast, other seed plant lineages (with the exception of Gnetales) have little evidence of polyploidy [66, 190]. As such, the major question is no longer "How many angiosperms are polyploid?", but rather "How many episodes of genome duplication have various lineages experienced?" and "Did specific genes resulting from genome duplication serve to fuel adaptive radiations?"

Despite its relatively high chromosome number of $2n = 26$, very few duplicate gene pairs were detected in *Amborella* based on the initial 8629 ESTs analyzed by Cui et al. [41], and no K_s signal for ancient genome duplication was observed. The Ancestral Angiosperm Genome Project (AAGP) has now produced more than 20800 conventional Sanger ESTs and 800000 ESTs generated via 454 sequencing technology (http://www.454.com/enabling-technology/index.asp) for *Amborella*. This K_s analysis identified 279 paralogous gene pairs with sequence length greater than 300 bp: more than four times the 69 pairs identified by Cui et al. [41] . Even with this greatly expanded EST data set, there is no evidence for an early genome duplication in *Amborella* (Fig. 6.2; AAGP, unpublished data). However, the lack of evidence for ancient polyploidy in *Amborella* K_s distributions does not preclude the possibility of ancient polyploidy in either this lineage or in the genome of an early angiosperm or angiosperm precursor. The absence of a polyploidy K_s signal in *Amborella* could also occur if evidence of ancient WGD has been eroded by gene death and/or satu-

ration of synonymous substitutions [41] to the point where it is now undetectable in analyses of EST samples. Alternatively, the earliest duplication peak detected in the *Nuphar* analysis (Fig. 6.2; discussed before) may trace back to an early genome duplication that occurred in the common ancestor of *Nuphar* and all extant angiosperm lineages other than *Amborella*. Complete sequencing of the *Amborella* nuclear genome is needed to test these alternative hypotheses [192].

Duplication of several MADS-box genes, which control floral organ identification and development, also suggests genome doubling early in angiosperm history or perhaps prior to crown angiosperm origins [30, 78, 87, 92, 93, 191, 197, 234]. For example, gymnosperms have only one B-function lineage, whereas all angiosperms have at least two such lineages (homologs of AP3 and PI). The two B-function gene lineages, which include homologs of AP3 and PI, respectively, appear to have originated via duplication of a single B-function gene at some point prior to the origin of the angiosperms [87]. An ancient duplication event of comparable timing occurred in the C-function lineage, forming two lineages in angiosperms, one with AG homologs (with roles in stamen and carpel identity), and the other with D function (with a role in ovule formation) [92]. Similarly, duplication of SEP genes resulted in the AGL2/3/4 (SEP1/2/4) and AGL9 (SEP3) lineages in the common ancestor of the angiosperms [234]. WGD is also favored by analysis of protein-protein interaction networks among MADS domain proteins [217]. Interestingly, the latter study suggests that heterodimerizing factors appear to derive from duplication of homodimerizing ancestors. Duplications of these key floral organ identity genes may have been important in the origin of the flower, but the exact timing of these gene duplications remains unclear. Given that some of these MADS-box gene duplications are thought to be quite old (over 260 million years, well before the origin of the crown angiosperms; see [87]) and others more recent, were there several genome-wide duplication events

along the stem lineage leading to the crown angiosperms? Similar questions also apply to early events in angiosperm diversification. Phylogenetic studies have revealed duplication of MADS-box genes *within* the basal eudicot clades; these genes include the AG, AP3, AP1 , and SEP gene lineages [78, 91–93, 117, 234]. Could these correspond to a genome-wide duplication event early in eudicot evolution (e.g., the γ event; Fig. 6.1)? MADS protein-protein interaction data imply successive duplications from angiosperm origins through the rosid-asterid split [217]. Forthcoming asterid genome sequences for *Mimulus* and one or more species of Solanaceae may help resolve this question.

Reconstructing ancestral genomes prior to genome duplication – Now that WGD is known to be both frequent and ubiquitous across angiosperm history, a major challenge is the reconstruction of the ancestral genomes of lineages prior to genome duplication. Reconstruction of the ancestral genomes of lineages prior to genome duplication improves our ability to resolve correlated gene arrangements among taxa [27], which in turn provide a valuable framework for inference of shared ancestry of genes and for the utilization of findings from model organisms to study less well-understood systems. Such alignments have also revealed patterns of differential gene loss following genome duplication, differential gene retention associated with evolution of some morphological complexity, and unexpectedly large variation among taxa in DNA substitution rates [207].

Such reconstructions are challenging, however, because whole-genome duplication may be followed, over evolutionary time, by genome downsizing (paralog extinction and the loss of noncoding DNA), rearrangement through intra- and interchromosomal movement of genetic material, and continual gene duplication events that do not involve WGD. However, reconstruction of ancestral genomes can potentially be achieved through comparative analysis of genome content and structure for extant

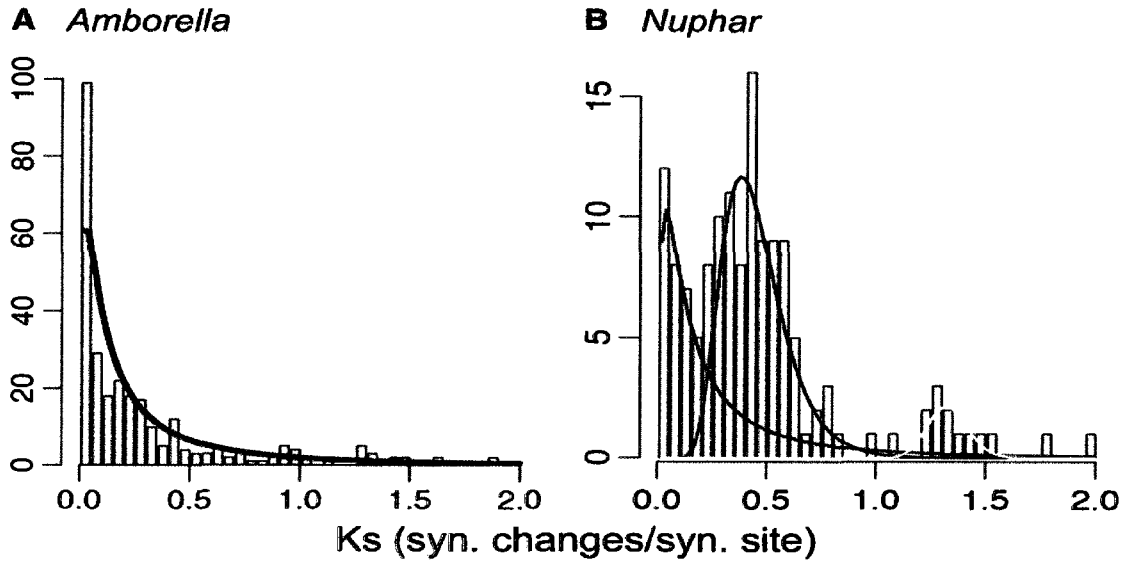

Figure 6.2: A greatly expanded *Amborella* expressed sequence tag (EST) data set provides much greater power for detecting ancient genome duplication events. (A) K_s analysis of *Amborella trichopoda* based on 28000 conventional ESTs (Sanger), and approximately 800000 454 ESTs; this facilitated the assembly of 701 gene pairs. Although this expanded *Amborella* EST data set provides much greater power to detect ancient polyploidy than was available for the estimation in [41], there is still no evidence for genome duplication based on this K_s analysis. (B) K_s analysis of 134 gene pairs identified from 8442 *Nuphar* ESTs [41] shows three components to the K_s plot, including a background duplication (green), and two significant older (one in black and one in yellow, Fig. 6.2) components representing likely ancient polyploidy events.

species, combining information from gene duplication histories (developed through analysis of sequence similarities or gene trees) and chromosomal gene locations to identify duplicated blocks of genes or DNA sequence dispersed among the chromosomes in one or more species. Sequence analysis tools are not directly applicable to this problem because all pairs of duplicates in doubled genomes were generated at the same historical moment. A rapid (linear-time) algorithm to find the ancestral genome that minimizes the genomic distance (minimum number of inversions and translocations) to a present- day genome has been available for some time [51]. Unfortunately, there are numerous rather different solutions that are equally parsimonious. A new procedure–guided genome halving–seeks to counteract this problem by guiding the reconstruction using one or more reference, or outgroup, genomes [174, 238, 241, 242]. The principle behind the guided genome halving approach is that each time a choice between two equally good construction steps is encountered by the halving algorithm, the outgroup information comes into play to see if one of the choices corresponds more to the structure of the reference genome(s).

Here we use the complete nuclear genome sequences of *Populus* (Salicaceae) and *Vitis* (Vitaceae) to attempt to reconstruct the ancestral genome of *Populus* prior to genome duplication (Fig. 6.3) as an example of the process of reconstructing ancestral genomes. Paralogous gene pairs in *Populus* and their single orthologs in *Vitis*, where these exist, were used as input. Both *Vitis* and *Populus* are rosids, with Vitaceae sister to all other rosids and Salicaceae a member of Malpighiales (in the eurosid I or Fabidae clade) [9, 34]. We scaled up the guided genome-halving algorithm to handle the very large number of genes analyzable in the *Populus* and *Vitis* data, introducing all pertinent information derivable from the outgroup (*Vitis*) into the halving algorithm as applied to *Populus*. The comparison reveals blocks of genes that have remained unaltered over the perhaps 109–124 million years since the two

lineages shared a common ancestor (Fig. 6.3) [193, 224]. A recent putative WGD event in *Vitis* [216] might be problematic in these comparisons, but there was no evidence of this in the analyses we conducted.

Many large, duplicated gene blocks observed in two large segments of the *Populus* genome match to one segment in the *Vitis* genome, attesting to the high degree of conserved synteny (gene order within blocks of sequence) and colinearity (ordering of the syntenic blocks) during divergence of the *Vitis* and *Populus* lineages, as well as retention of duplicate blocks during the diploidization of the ancestral *Populus* tetraploid (Fig. 6.3). Nevertheless, details of some of the reconstructed chromosomes cannot be assigned a high degree of confidence. For example, much of chromosome 7 consists of short syntenic blocks concatenated by the program based on a "minimum inversions plus translocations" criterion. This juxtaposes genes that are in fact nowhere adjacent in the *Vitis* and *Populus* data. Improvements to existing algorithms, now underway, will exploit new evidence from additional related genomes.

6.3 Have ancient polyploidy events resulted in increased species richness?

Given the many purported benefits of polyploidy [108–110] and the suggested relationship of genome duplication to speciation [123, 124, 223], rates of diversification may be higher in polyploid lineages than in diploid groups (due either to increased rates of speciation, decreased rates of extinction, or both). The identification of ancient WGD events at many points in angiosperm phylogeny provides the opportunity to assess the correspondence between inferred genome duplication events and large diversifications, and hence the role of polyploidy in "macro-diversification."

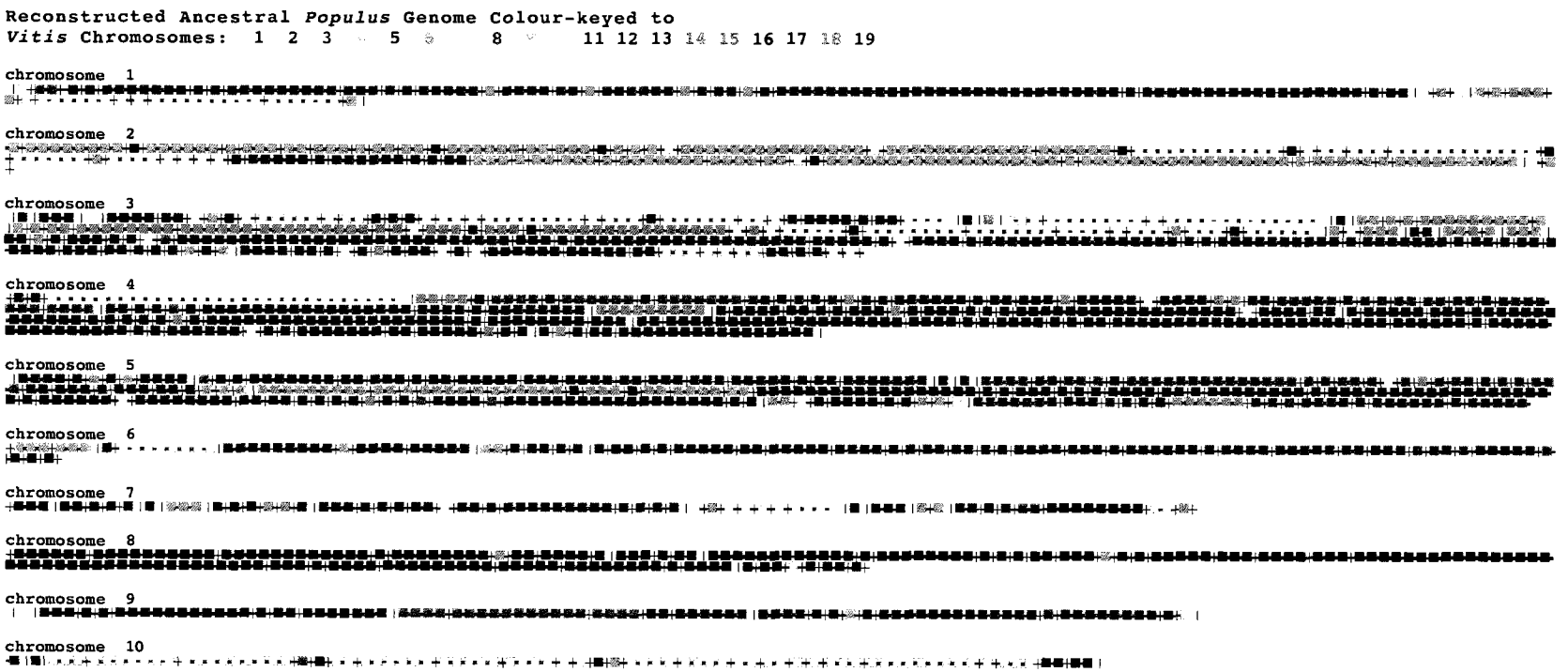

Figure 6.3: Reconstruction of the preduplication ancestor of *Populus*. The data contain 2940 genes with two copies in *Populus* and at most one copy in *Vitis*. Of these, 211 "singleton" genes were excluded from the analysis: no two of copies of a singleton have an immediately neighboring gene, an "adjacency", in common. The reconstruction algorithm finds the genome that minimizes the genomic distance, i.e., the number of rearrangements (inversions and translocation), from the ancestral tetraploid to the modern *Populus* genome, while simultaneously minimizing the distance from the diploid version of the ancestor to *Vitis*. The 10 reconstructed chromosomes, containing only the 2020 genes on *Vitis* chromosomes, are shown in wrapped form. Added after the reconstruction: adjacencies present in the original data three times, i.e., twice in *Populus* and once in *Vitis*, are indicated by triple dashes (≡), those present twice by double dashes (=), those present once by single dashes (−), and those for which there is only indirect evidence from the rearrangement analysis by a blank space. Intrachromosomal breakpoints in comparison with one of the *Populus* copies indicated by vertical.

Anecdotal data suggest that polyploid lineages are "successful," but a statistical association of polyploidy and species richness has not been rigorously tested. To address this question, we compared species richness in clades that are ancient polyploids with sister clades that are not. The overall diversification rate (r) for angiosperms was estimated based on the methods described in [126] . Because the estimation of this parameter is contingent on the rate of extinction (ε), which is an unknown, we estimated r across a range of extinction rates [8]. We also calculated values of r over a range of plausible age estimates for crown group angiosperms (132 Mya), based on first unambiguous occurrence in the fossil record [74]; 170 Mya based on recent molecular age estimations [138,224]; 250 Mya based on other estimates based on molecular sequence data [127,168]. Next we calculated the probability of observing the extant number of species in several putative clades of polyploid origin given these estimated global rates for angiosperms, conditioned on the assumed age of the crown group (e.g., r_{G-132}, see Table 6.1). All calculations were done using GEIGER 1.0–91 [72].

Current standing diversity of each polyploid clade was tabulated based on values taken from Stevens [204], as well as Olmstead et al. [145], and Wojciechowski et al. [226]. Absolute ages of groups were taken from [224]. We selected these ages because they represent estimates from the most comprehensive taxon sampling to date. However, taxon sampling by Wikström et al. was not always dense enough across clades of interest (Figs. 6.4,6.5,6.6,6.7) to pinpoint potential duplication events. In these instances, we used the age of the next most inclusive node that would include the inferred duplication event (e.g., age of Fabaceae, when the duplication event occurred after the origin of the crown group). By doing this, we provided a conservative estimate of diversification rates for each of the clades.

Successfully conducting this exercise is difficult in that it requires both reliable genomic data for ancient polyploidy, plus a sufficient breadth of data to be able to

Clade	No. extant species	Age[a]	$\varepsilon = 0.0$ $r_{G-132} = 0.03690$	$\varepsilon = 0.3$ $r_{G-132} = 0.03619$	$\varepsilon = 0.5$ $r_{G-132} = 0.03475$	$\varepsilon = 0.7$ $r_{G-132} = 0.03186$	$\varepsilon = 0.9$ $r_{G-132} = 0.02457$
Brassicaceae	3710	22	$\ll$ **0.001**	$\ll$ **0.001**	$\ll$ **0.001**	$\ll$ **0.001**	$\ll$ **0.001**
Cleomaceae	300	22	$\ll$ **0.001**	$\ll$ **0.001**	**0.002732658**	**0.01243294**	0.0866002
Within Fabaceae–	7000	63	$\ll$ **0.001**	$\ll$ **0.001**	$\ll$ **0.001**	$\ll$ **0.001**	$\ll$ **0.001**
Core Poaceae	10000	15	$\ll$ **0.001**	$\ll$ **0.001**	$\ll$ **0.001**	$\ll$ **0.001**	$\ll$ **0.001**
Solanoideae	1925	39	$\ll$ **0.001**	$\ll$ **0.001**	$\ll$ **0.001**	$\ll$ **0.001**	$\ll$ **0.001**
			$\varepsilon = 0.0$ $r_{G-170} = 0.02866$	$\varepsilon = 0.3$ $r_{G-170} = 0.02811$	$\varepsilon = 0.5$ $r_{G-170} = 0.02698$	$\varepsilon = 0.7$ $r_{G-170} = 0.02475$	$\varepsilon = 0.9$ $r_{G-170} = 0.01908$
Brassicaceae	3710	22	$\ll$ **0.001**	$\ll$ **0.001**	$\ll$ **0.001**	$\ll$ **0.001**	$\ll$ **0.001**
Cleomaceae	300	22	$\ll$ **0.001**	$\ll$ **0.001**	$\ll$ **0.001**	**0.002709716**	**0.03925442**
Within Fabaceae–	7000	63	$\ll$ **0.001**	$\ll$ **0.001**	$\ll$ **0.001**	$\ll$ **0.001**	$\ll$ **0.001**
Core Poaceae	10000	15	$\ll$ **0.001**	$\ll$ **0.001**	$\ll$ **0.001**	$\ll$ **0.001**	$\ll$ **0.001**
Solanoideae	1925	39	$\ll$ **0.001**	$\ll$ **0.001**	$\ll$ **0.001**	$\ll$ **0.001**	$\ll$ **0.001**
			$\varepsilon = 0.0$ $r_{G-250} = 0.01948$	$\varepsilon = 0.3$ $r_{G-250} = 0.01911$	$\varepsilon = 0.5$ $r_{G-250} = 0.01835$	$\varepsilon = 0.7$ $r_{G-250} = 0.01683$	$\varepsilon = 0.9$ $r_{G-250} = 0.01297$
Brassicaceae	3710	22	$\ll$ **0.001**	$\ll$ **0.001**	$\ll$ **0.001**	$\ll$ **0.001**	$\ll$ **0.001**
Cleomaceae	300	22	$\ll$ **0.001**	$\ll$ **0.001**	$\ll$ **0.001**	$\ll$ **0.001**	**0.008363369**
Within Fabaceae–	7000	63	$\ll$ **0.001**	$\ll$ **0.001**	$\ll$ **0.001**	$\ll$ **0.001**	$\ll$ **0.001**
Core Poaceae	10000	15	$\ll$ **0.001**	$\ll$ **0.001**	$\ll$ **0.001**	$\ll$ **0.001**	$\ll$ **0.001**
Solanoideae	1925	39	$\ll$ **0.001**	$\ll$ **0.001**	$\ll$ **0.001**	$\ll$ **0.001**	$\ll$ **0.001**

Table 6.1: Significant departure in diversification rates of the major plant clades from the global diversification rate estimates. Boldfaced P values indicate significantly higher species diversity than expected under the global rate (r_{G-132}) of angiosperms based on a minimum age of 132 Myr, a global rate (r_{G-170}) based on a mean age of 170 Myr [138, 224], a global rate (r_{G-250}) based on a maximum age of 250 Myr [127, 168], and a current standing diversity of 261750 species ε is the extinction rate.

[a]Based on estimates of [224]

place the WGDs with confidence on a phylogenetic tree. Unfortunately, for most putative examples here, the exact placement of the genome duplication events proposed is not yet clear. Although the data and analyses presented certainly suggest that ancient polyploidy has led to a dramatic increase in species richness in several clades, until additional genomic studies are conducted to pinpoint the placement of those WGD events, the results should be considered preliminary. Nonetheless, in all but a few cases (calculations involving Cleomaceae), p-values were found to be highly signifi cant ($\ll 0.001$; Table 6.1), suggesting exceptionally high diversification rates in these polyploid clades.

Poaceae–The estimated age of a genome duplication event characterizing Poaceae (70- 50 million years before present: [25,128,151,178] is not only close to the molecular estimates for the age of the family (ca. 89 and 83 million years for stem and crown group Poaceae, respectively) [28, 84], but also similar to the age of grass fossils (55 million years) [40] assignable to the PACCMAD (Panicoideae, Arundinoideae, Centothecoideae, Chloridoideae, Micrairoideae, Aristidoideae, Danthonioideae) and BEP (Bambusoideae, Ehrhartoideae, Pooideae) clades [68]. The precise phylogenetic placement of the genome duplication is still unclear (as represented by angled line; Fig. 4), but if this event coincided with, or was close to the origin of core Poaceae (indicated by the slash mark on the tree), it would agree with the hypothesis that polyploidy promotes speciation. Core Poaceae are a very large group (658 genera and 9998 species), whereas the three early-branching subclades of Poaceae (Anomochlooideae, Pharoideae, Puellioideae) are all very small, as are the three families most closely related to Poaceae–Flagellariaceae, Joinvilleaceae, and Ecdeicoleaceae [129] (Fig. 6.4).

Brassicaceae and Cleomaceae–Recent data indicate that a genome-wide triplication occurred in Cleomaceae independently of WGD events in Brassicaceae [179]. Strengthening support for the occurrence of these doubling events well within Brassi-

cales is the recent genome sequence for *Carica* of Brassicales, which shows no evidence of genome duplications more recent than γ (discussed earlier [135]). Hence, the data suggest independent WGDs occurred in both Cleomaceae and Brassicaceae (Fig. 6.5). Virtually all families of Brassicales are very small–most have fewer than 15 species. In contrast, Capparaceae (480 species), Cleomaceae (300 species), and particularly Brassicaceae (3710 species) are all species-rich. Genome-scale data are not yet available for Capparaceae, but evidence for genome duplication in Cleomaceae and Brassicaceae is suggestive of an association between polyploidy and diversification.

Fabaceae–A genome-wide duplication event within the legumes has been documented and is postulated to have occurred immediately prior to the radiation that yielded approximately 7000 species [33,99,156]. However, the placement of this duplication event is much more problematic than those reviewed here. The short branches along the backbone of the tree for the Fabaceae, particularly in the papilionoids, makes placing this duplication and determining the sister group of the clade having the duplication diffi cult. Because the radiation of the family was so rapid, the uncertainty of the timing of the polyploid event puts just about every node within reach of the event (J. Doyle, Cornell University, personal communication). Considering the placement of the WGD in the common ancestor of *Hologalegina* and phaseoloids [156] (J. Doyle, personal communication), the baphioids, a clade of only five genera and fewer than 20 species, are the immediate sister to the remaining papilionoids, a clade of 7000 species. If this placement is correct, it would again favor the hypothesis that genome duplication may be associated with diversification (Fig. 6.6).

Solanaceae–A WGD event in Solanaceae has been dated from ca. 50 - 52 million years before present [178]. Bell et al. [15] estimated that the age of stem Solanaceae is approximately $49 - 68$ Myr and the crown group $32 - 50$ Myr, in agreement with the estimated age of the duplication as having occurred within Solanaceae. This WGD

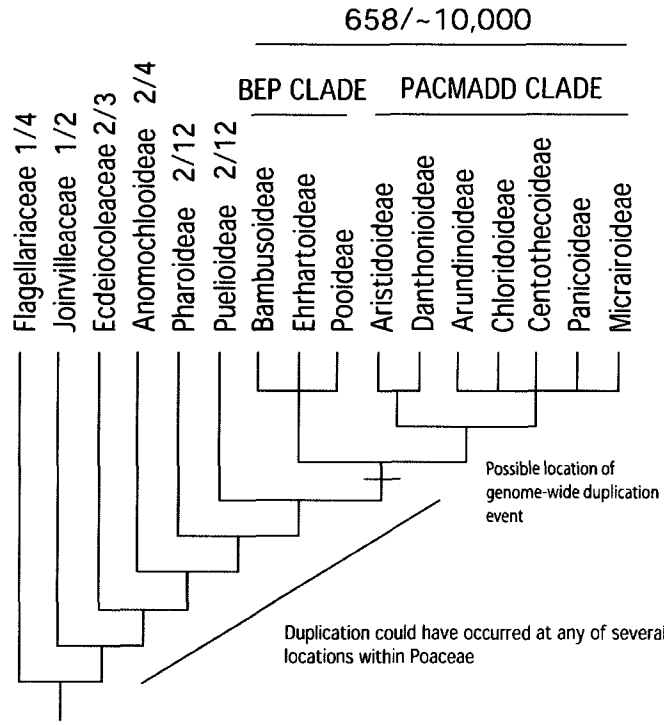

Figure 6.4: Placement of inferred genome duplication event detected in Poaceae [151]; summary tree modified from that of the Grass Phylogeny Working Group [68]. The precise phylogenetic placement of the genome duplication is still unclear (as represented by angled line), but if this event coincided with, or was close to the origin of core Poaceae (indicated by the slash mark on the tree), it would agree with the hypothesis that polyploidy promotes speciation. The numbers beside each taxon are the number of genera/species.

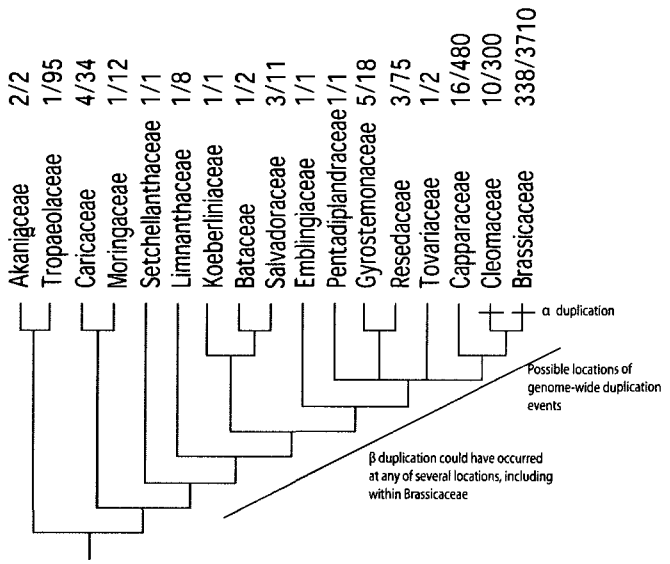

Figure 6.5: Placement of inferred genome duplication events in Brassicales–the α and β events occurred in Brassicales, with the α event possibly occurring early in the diversification of Brassicaceae (indicated by slash mark). The precise phylogenetic placement of the β genome duplication is unclear (as represented by angled line), but β likely occurred after the divergence of Caricaceae (see text). An additional triplication event (indicated by slash mark) took place in Cleomaceae [179]; summary tree modified from D. Soltis et al. [190] and the Angiosperm Phylogeny Website [204]. The numbers beside each taxon are the number of genera/species.

event characterizes either the clade corresponding to Solanoideae, or perhaps more likely, what has been referred to as the $x = 12$ clade [145] (Fig. 6.7). The $x = 12$ clade comprises the Solanoideae and Nicotianoideae. Members of Solanaceae outside of the $x = 12$ clade have a range of chromosome numbers, most from $x = 7 - 11$, but there are no relic $x = 6$ lineages that could be potential parents of the $x = 12$ clade [145]. Hence, our results are consistent with a WGD event that occurred in the ancestor of the $x = 12$ clade or Solanoideae clade (Fig. 6.7). However, as with the other examples noted, more precision is needed in ascertaining the placement of this ancient duplication.

If this duplication event occurred in either the ancestor of the $x = 12$ clade or the

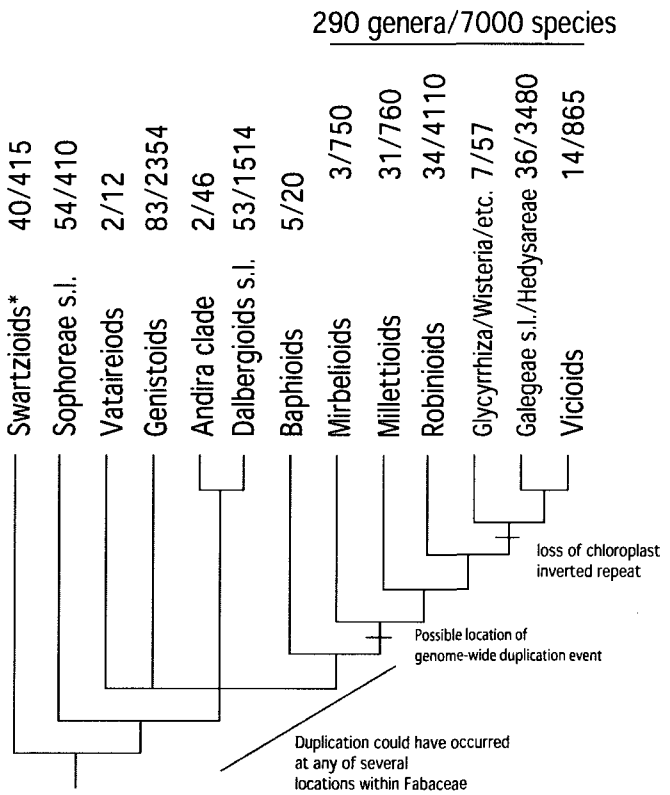

Figure 6.6: Putative placement of inferred genome duplication event in Fabaceae [99,156]. The precise phylogenetic placement of the genome duplication is still unclear (as represented by angled line), but one strong possibility for the timing of this event is indicated by the slash mark on the tree; the location of the loss of the chloroplast inverted repeat is also indicted by a slash mark; summary tree modified from Wojciechowski et al. (2004) [226] . The numbers beside each taxon are the number of genera/species

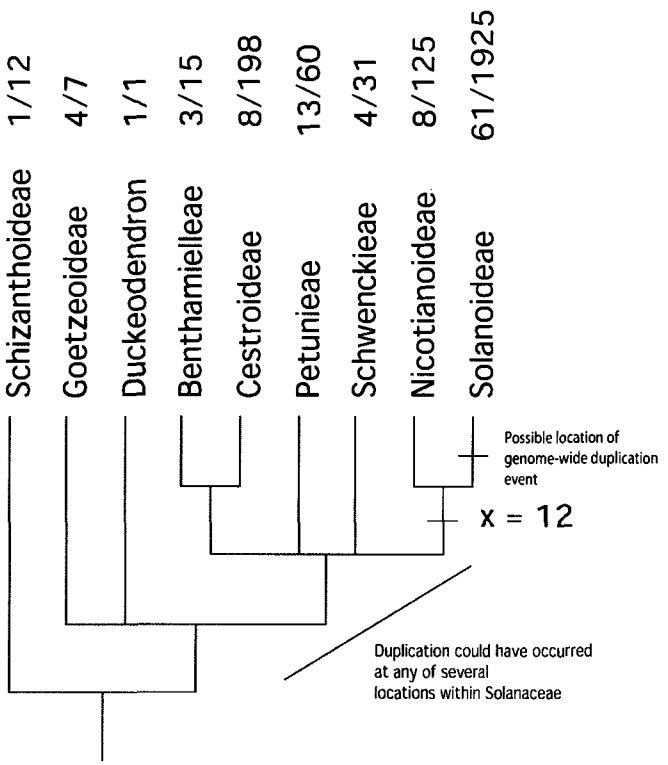

Figure 6.7: Putative placement of inferred genome duplication event in Solanaceae [178]. The precise phylogenetic placement of the genome duplication is still unclear (as represented by angled line), but one possibility is that this event coincided with the origin of Solanoideae (indicated by the slash mark on the tree); summary tree modifi ed from Olmstead et al. (in press) . $X = 12$ refers to the $X = 12$ clade of Olmstead et al. [145]. The numbers beside each taxon are the number of genera/species.

Solanoideae, this ancient WGD event likely had a major impact on diversifi cation. Solanoideae is by far the largest clade within Solanaceae, comprising 61 genera and 1925 species [204]; other clades in Solanaceae are considerably smaller.

6.4 Additional evidence and conclusions

Clues regarding the importance of polyploidy in angiosperm diversification may come from studies of recent radiations. Island floras, particularly those inhabiting volcanic chains, may form microcosms for understanding processes behind largerscale diversifi cations. Dispersal to volcanic archipelagoes involves the colonization of newly formed habitats that change dramatically over relatively short periods of time with island subsidence. This cycle of environmental instability can repeat itself when further dispersal occurs to younger islands as they form. Early angiosperms may have experienced similar environmental instabilities after the Cretaceous-Tertiary (K-T) Boundary [132].

The Hawaiian flora has the highest incidence of polyploidy known, and most Hawaiian species are paleopolyploid, having evolved polyploidy prior to the dispersal of their ancestors to Hawaii [36]. Both polyploid hybrid and allopolyploid Hawaiian radiations have been studied in detail. The species-rich and morphologically diverse Hawaiian endemic mints, which are high polyploids ($2n = 64, 66$), descend from a relatively recent hybridization event between two different polyploid North American lineages, one with bird-pollinated flowers, the other with insect-pollinated ones [116]. Their morphological and ecological variation is extensive; plants range from subalpine vines to rainforest shrubs, flowers may have either bird- or insect- pollinated anatomies, and seed dispersal patterns may depend on either dry or fl eshy fruits. In another example, it has been shown that the Hawaiian silverswords, with their incred-

ible vegetative diversity, are recent allopolyploids, and the presence of two divergent genomes in their colonizing ancestor may have helped promote adaptive radiation in the alliance [14]. Molecular studies of polyploid genome formation have indicated that phenomena such as substantial intragenomic rearrangement and altered gene regulatory relationships, and in allopolyploids, fixed heterozygosity, can contribute to evolutionary flexibility [108–110,196,209,221,222]. Polyploid radiations in modern unstable island environments may therefore provide clues as to why some angiosperm lineages (e.g., Poaceae and Fabales) underwent high diversification long after their first appearance in the fossil record.

Nowhere else in the history of life is the influence of polyploidy more apparent as a possible diversifying force as it is among the angiosperms. Only the tetraploidization event in the stem lineage of ray-finned fish may come close, having potentially resulted in $>$ 20000 living species [69, 101]. Despite clear evidence for WGD underlying all vertebrates [140], many fewer lineage-specific polyploidy events have been detected than among angiosperms, which have a far shorter combined evolutionary history. Much of this disparity can be attributed to the different nature of plant vs. animal development vis-à-vis compensation for gene dosage [66, 125]. Subfunctionalization of gene duplicates has been shown to be a prominent molecular evolutionary force for coping with WGD in both flowering plants and rayfinned fish [2, 4, 228]; still, the angiosperms have diverged into many more habitats, life forms, and biotic interactions, and as such, the importance of their polyploid heritage cannot be underestimated as a force of truly global impact. Future work addressing the influence of polyploidy from the population to whole-lineage levels will clearly be important for understanding these phenomena.

Chapter 7

Gene loss under neighbourhood selection following whole genome duplication and the reconstruction of the ancestral populus genome

Zheng C, Wall PK, Leebens-Mack J, dePamphilis C, Albert VA, Sankoff D. 2008. Journal of Bioinformatics and Computational Biology, in press. [Conference version: Zheng C, Wall PK, Leebens-Mack J, dePamphilis C, Albert VA, Sankoff D. 2008.The effect of massive gene loss following whole genome duplication on the algorithmic reconstruction of the ancestral Populus diploid. Proceedings of CSB 2008, 261-271.] I was responsible for designing the algorithm and getting all of the results in this work. I also collaborated fully in writing and preparing the manuscript, especially technical write-ups and graphics. P.K. Wall, a member of C. dePamphilis' research

team, provided the *Populus* and *Vitis* data organized into homology sets. Some of the work and manuscript revision was done in consultation with V.A. Albert, and J. Leebens-Mack also provided helpful information.

Abstract

We develop criteria to detect neighbourhood selection effects on gene loss following whole genome duplication, and apply them to the recently sequenced poplar (*Populus trichocarpa*) genome. We improve on guided genome halving algorithms so that several thousand gene sets, each containing two paralogs in the descendant T of the doubling event and their single ortholog from an undoubled reference genome R, can be analyzed to reconstruct the ancestor A of T at the time of doubling. At the same time, large numbers of defective gene sets, either missing one paralog from T or missing their ortholog in R, may be incorporated into the analysis in a consistent way. We apply this genomic rearrangement distance-based approach to the poplar and grapevine (*Vitis vinifera*) genomes, as T and R respectively. We find that the loss of paralogs from T introduces far more noise into the reconstruction of A than does the loss of the ortholog from R.

7.1 Introduction

Following an episode of whole genome doubling (WGD), gene duplicates are lost at a high rate through processes such as pseudogenization and deletion of chromosomal segments containing one or more genes, while intra- and interchromosomal rearrangement mechanisms redistribute chromosomal segments both large and small across the genome. The genome of the present-day descendant can be largely decomposed into a set of duplicated DNA segments dispersed among the chromosomes, with all the duplicate pairs exhibiting a similar degree of sequence divergence, and with single-copy segments interspersed among them. In this paper, we introduce approaches to analyzing the evolution of doubled genomes, based entirely on gene order evidence, in order to explain aspects of the gene loss process and to reconstruct the rearrangement steps leading from the doubled ancestral genome to the present day descendant.

Though syntenic evidence, namely duplicated segments containing several genes in corresponding order, have long been used as evidence for WGD, studies of duplicate gene loss have focused on functional changes and divergence rates within individual gene families, to the exclusion of gene order considerations. Here we investigate how the fate of duplicate genes is correlated, or not, to the retention or loss of nearby genes on the same chromosomes.

As for reconstructing rearrangement history, a linear-time "genome halving" algorithm, based only on the ordering of duplicated chromosomal segments, can find an ancestral genome that minimizes the genomic distance to the present-day genome [50, 51]. This does not suffice, however, as a solution to the reconstruction problem, since there may be a large number of very different, equally optimal solutions. Here we use a guided genome halving (GGH) strategy to overcome this non-uniqueness, guiding the reconstruction of the ancestor by one or more reference,

or outgroup, genomes. This strategy does not sacrifice the optimality of the halving solution.

The flowering plants are well-known for numerous historical events of genome doubling [41]. The recently sequenced poplar genome (*Populus trichocarpa*) [213], which shows very clear evidence of genome duplication 60 - 65 million years ago, and the grapevine genome (*Vitis vinifera*) [81, 216], whose ancestor diverged before the aforementioned duplication, provide a pair of analytical incentives to the GGH strategy. On the one hand, the poplar data has an order of magnitude more duplicated elements than have previously been analyzed, straining computational resources. On the other hand, the richness of the data allows us to assess neighbourhood selection effects on duplicate gene loss and the implications of this loss of genes from thousands of duplicated pairs on the accuracy of ancestral genome reconstruction.

This paper thus contributes three advances on the methodological level: first, a way of analyzing chromosomal neighbourhood selection effects on the retention or loss of duplicated genes, second, the scaling up, by more than an order of magnitude, of the amount of data amenable to our GGH analysis, and third, the incorporation into GGH of data from gene duplicate pairs that have lost one member, making use of chromosomal context in both the genome that can be traced to the doubling event and in the outgroup.

7.1.1 Outline

In Section 7.2, we describe the sources for our data and how we processed them to obtain the gene sets for the selection study and the GGH analysis. In Section 7.3, we present our method and results for detection of neighbourhoood selection effects. In Section 7.4 we sketch the necessary background about genomic rearrangement

distance and the genome halving and GGH algorithms. In Section 7.5 we present the GGH algorithm incorporating both full and defective gene sets. We apply this method to the full gene sets in combination with one or both of two defective gene sets from *Populus* and *Vitis* in Section 7.6. We present the reconstructed undoubled *Populus* ancestor based on over 6000 gene sets and evaluate the evolutionary signal versus noise (a) in the ancestor-*Populus* and ancestor-*Vitis* comparisons, (b) in the full and defective gene sets, and (c) in genes with two or three common adjacencies in the data and those with weaker positional evidence.

7.2 The *Populus-Vitis* comparison

Annotations for the *Populus* and *Vitis* genomes were obtained from databases maintained by the U.S. Department of Energy's Joint Genome Institute [213] and the French National Sequencing Center, Genoscope [81], respectively. An all-against-all BLASTP search was run on a data set including all *Populus* and *Vitis* protein coding genes, and orthoMCL [113] was used to construct 2104 full and 4040 defective gene sets, in the first case containing two poplar paralogs (genome T) and one grape ortholog (genome R), and in the second case missing a copy from either T or R. The chromosomal location and orientation of these paralogs and orthologs was used to construct our database of gene orders for these genomes, and the input to the GGH algorithm. In addition, 740 *Populus* single-copy genes with known chromosomal location but with no orthology detected in *Vitis* were added for the study of neighbourhood selection effects.

7.3 Neighbourhood effects

The data include 3944 single-copy genes with or without orthologs in *Vitis*. We assume that the single-copy status of most of these genes is simply a consequence of the loss of their duplicates following WGD, either through pseudogenization, outright deletion or other process, though of course there may be a small proportion with other explanations.

We may ask how the process of gene loss are distributed throughout the genome. In particular,

- is there any spatial non-randomness in the choice of duplicate pairs that will become single-copy? Are such pairs clustered together or randomly spaced in the genome?
- for any cluster of neighbouring gene pairs that becomes single-copy, do the remaining copies tend to reside on the same chromosome, or does each pair of duplicates "decide" independently from its neighbours which chromosome keeps the gene and which loses it?

7.3.1 Choice of pairs to become single-copy.

One difficulty in addressing these problems is that we do not know which genes in the present-day genome are neighbours because they have always been neighbours and which have become neighbours recently, for example through chromosomal rearrangement operations, and this is especially problematic for single-copy genes.

To circumvent this problem, we consider only sets of single-copy genes bounded at both ends by a pair of duplicate genes on the same two chromosomes, as in Figure 7.1. Such a configuration, which we take as our basic unit of analysis, can arise

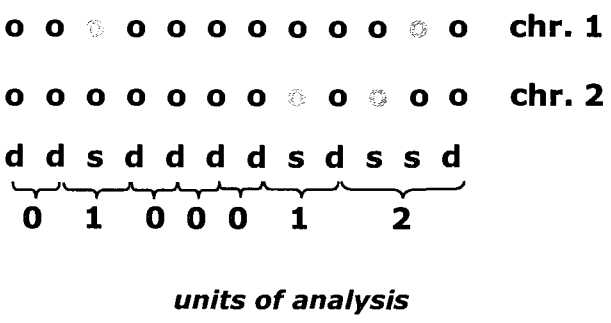

Figure 7.1: Analytical units for the study of single-copy genes. Black circles represent existing genes, gray circles represent hypothesized lost genes. d = duplicate gene pair, s= single-copy remaining. Numbers indicate length of string in which one member of each pair had been lost.

only infrequently by rearrangement, especially if the genome contains more than a handful of different chromosomes. The unlikeliness of these units arising by chance rearrangement explains why it suffices to take only one flanking pair of duplicates at either end, and why we need not be more stringent and ask for two consecutive such pairs at either end, with the consequent loss of valuable data for our analysis. We assume that the single-copy status of all the intervening genes, on one or the other of the two chromosomes, arose through the loss of one copy from a corresponding position on the other chromosome. Note that these sets of single-copy genes are not "paralogons", since the genes are distributed on two chromosomes, although by hypothesis they are descended from two paralogons, duplicate regions on the two chromosomes.

Thus the statistic we investigate is the frequency of the number of single-copy genes, on one chromosome or the other, between any two consecutive pairs of duplicate genes. Were the process of gene loss independent across two chromosomes of an artificial genome consisting of all the currently conserved duplicated genes (counting a pair only once if it is both at the end of one analytical unit and the beginning of another, e.g., the penultimate d in Figure 7.1), plus two copies of the currently

single-copy genes, we can predict this frequency distribution. Let n be the total number of duplicated pairs in this artificial construct. Then after S random selections of pairs from $\{1,\ldots,n\}$, with replacement but with only the first selection of any one pair resulting in a conversion to single-copy status, the probability that any particular duplicate pair A be conserved is $(1-\frac{1}{n})^S$ and the probability that the run of m pairs following A have been converted to single copy status, followed by a conserved duplicate pair B, is thus $(1-\frac{1}{n})^{2S}\left(1-(1-\frac{1}{n})^S\right)^m$, as long as A is in the first $n-(m+1)$ positions in the genome.

Then $\mathrm{E}[f(m)]$, the expected number of single-copy strings of length m, after S samples is

$$(n-m-1)(1-\frac{1}{n})^{2S}\left(1-(1-\frac{1}{n})^S\right)^m \tag{7.3.1}$$

Requiring the predicted proportion of single-copy genes $1-(1-\frac{1}{n})^S$ to be equal to the observed proportion $\frac{r}{n}$, leads to the predicted frequency

$$(n-m-1)\left(1-\frac{r}{n}\right)^2\left(\frac{r}{n}\right)^m. \tag{7.3.2}$$

Comparing the predicted versus the observed frequency of single-copy unit of various sizes, as in Figure 7.2 makes it clear that the null hypothesis of random choice of which duplicate pairs to convert to single-copy status would be difficult to dispute.

7.3.2 Concentration of single-copies on one chromosome.

There remains, however, the second type of neighbourhood selection we have mentioned, namely once a pair of duplicates is to change to single-copy status, is there an influence from neighbouring genes on which copy is to be lost and which to be

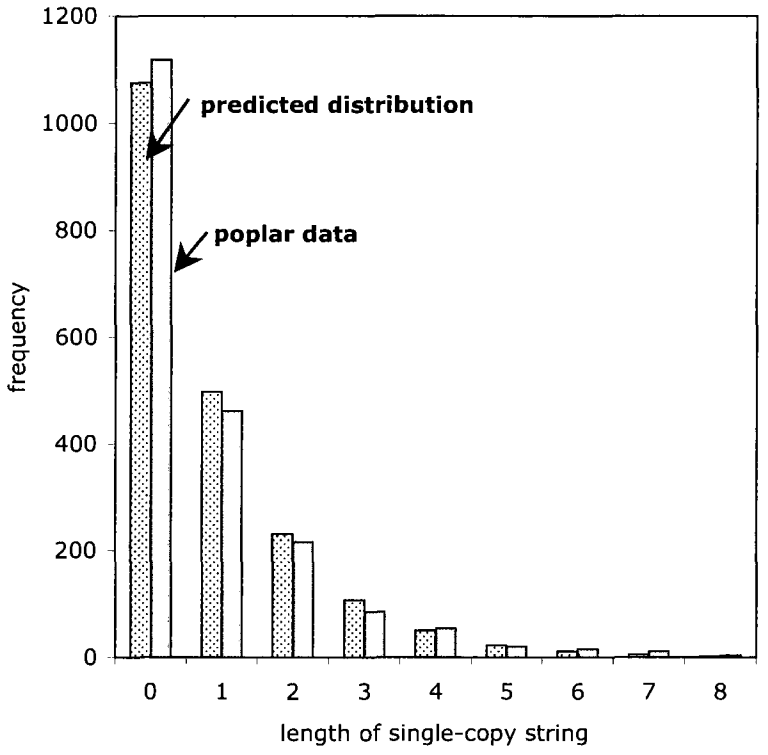

Figure 7.2: Congruence of predicted versus observed lengths of strings of single-copy genes.

retained? Under the hypothesis of no neighbour effects, configurations like those on the right of Figure 7.3, where both retained copies are on the same chromosome, should occur half of the time. Similarly for all m, we can calculate the probability that exactly q out of m genes occur on the same chromosome and $m-q$ on the other is $B(m,q)+B(m,m-q)$, for $q=0,\ldots,\lfloor\frac{m}{2}\rfloor$, where B is the binomial distribution.

The cumulative probability that q or fewer single-copy genes, out of m, will appear on the same chromosome is $\Pr_m(q) = \sum_{i=0}^{q} B(m,i) + B(m-q)$, for $q = 0,\ldots,\lfloor\frac{m}{2}\rfloor$. Plotting $\mathrm{Fr}_m(q)$, the observed cumulative relative frequency of q against this prediction $\Pr_m(q)$ should help us decide if the null hypothesis of randomness is justified. We use the cumulative distributions rather than the probability and relative frequency distributions themselves in order to combine the results for different

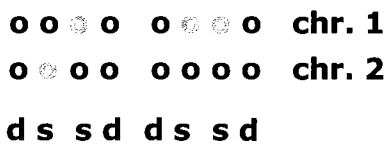

Figure 7.3: Single-copy gene evidence for neighbourhood selection. Black circles represent existing genes, gray circles represent hypothesized lost genes. d = duplicate gene pair, s= single-copy remaining. In the unit on the right, neighbours are conserved, on the left one gene is dropped from each chromosome.

values of m. Thus, in Figure 7.4 we superimpose plots of $\mathrm{Fr}_m(q)$ against $\mathrm{Pr}_m(q)$, for $m = 2, \ldots, 8$ and for $q = 0, \ldots, \lfloor \frac{m}{2} \rfloor$. There should be 23 points on such a plot, but seven of them are necessarily coincidental at (1,1).

The plausible alternative hypotheses are a widespread neighbourhood selection effect, possibly at the transcriptional level, perhaps involving co-regulation or common regulatory elements, or deletion-based effect whereby gene loss involves a chromosomal segment affecting more than one gene at a time. The latter is perhaps the simplest and most obvious choice, unless pseudogenization is the usual avenue to gene loss, and has been proposed in the context of gene loss after WGD in yeast [214], though selection-based explanations have been offered by other authors [32] and our own method (work to be reported elsewhere) yields even stronger support for selection rather than random large deletions than the poplar work reported here. Under the null hypothesis, there is no reason for there to be more points above the diagonal than below it, and the particular concentration of points where $\mathrm{Fr}_m(q)$ is much greater than $\mathrm{Pr}_m(q)$ and $\mathrm{Pr}_m(q) < \frac{1}{3}$ suggests that the low values of q, where almost all the single-copy genes in each analytical unit are on the same chromosome, are greatly over-represented in these data.

With the poplar data, were random deletions involving segments larger than one gene a quantitatively important explanation for the results in Figure 7.4, this

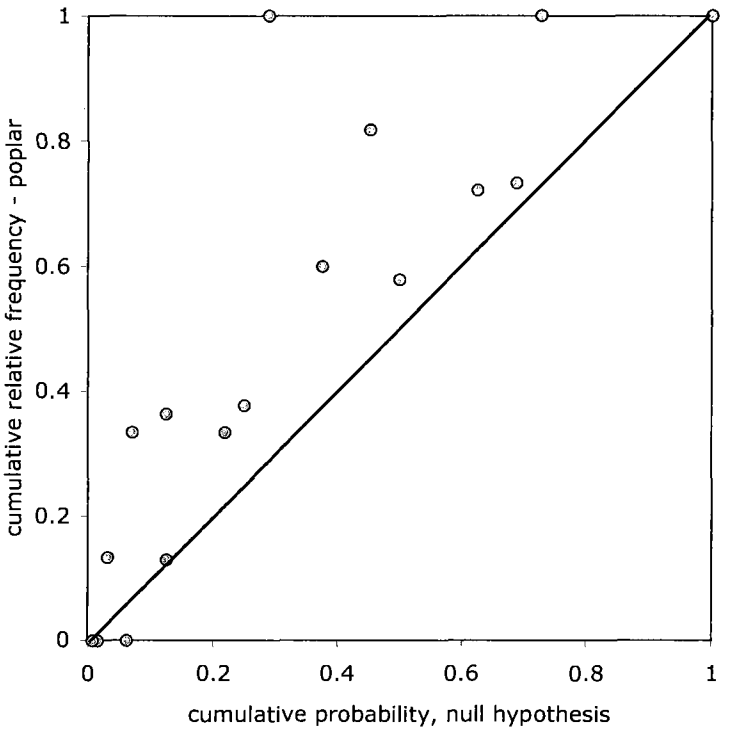

Figure 7.4: Observed cumulative relative frequency against predicted values, for q or fewer single-copy genes, out of m, to appear on the same chromosome, for $m = 2, \ldots, 8$ and for $q = 0, \ldots, \lfloor \frac{m}{2} \rfloor$. There should be 23 points on such a plot, but seven of them are necessarily coincidental at (1,1)

would necessarily also show up as deviation from the smooth geometric distribution in Figure 7.2, with a depressed value for $m = 1$ relative to increased values for at least $m = 2$ or higher m. This is emphatically not the case, and we can thus reject the large deletions theory in favour of a neighbourhood selection effect.

7.3.3 Long strings of single copies.

Along with the data summarized in Figures 7.3 and 7.4, we also found four analytical units with relatively long strings of single copies. These occurred in analytical units on chromosome pair (6,18), with 12 genes on chromosome 18 and only 2 on

chromosome 6, chromosome pair (2,14), with 12 genes on chromosome 14 and only 3 on chromosome 2, chromosome pair (13,19), with 15 genes on chromosome 13 and none on chromosome 19, and chromosome pair (8,10), with 16 genes on chromosome 10 and only 1 on chromosome 8.

These data provide additional striking confirmation of the tendency for single-copy genes to be largely on the same chromosome. The occasional instance of an active gene on the other chromosome is validation of our operational definition of an analytical unit: the surviving duplicate pairs delimiting either end of the single-copy segment did not get there through coincidental rearrangements, rather they define what remains of a genuine duplicated segment.

The extraordinary length of each of these four strings of single-copy genes remains unexplained, though functional selection effect are strongly suggested in several cases. For example, in the (2,14) pair and the (13,19) pair, about half of the genes also appear in a (single) cluster in the *Arabidopsis* genome. The genes in the (8,10) pair are almost evenly split between two such clusters in *Arabidopsis*. This suggests that the genes in question have a function-driven tendency to group together in a single-copy segment, a tendency which has survived separate WGD events in both *Populus* and *Arabidopsis*. This idea has recently been explored for yeast [158].

Moreover GO functional annotations show that four of the genes on chromosome 10 have nucleic acid or nucleotide binding capacity and three of the genes on chromosome 14 have hydrolase activity, in both cases more than could be expected by coincidence, and suggestive of some concerted functional connection favoured by the close linkage of these genes.

7.3.4 The units of analysis

It is only in the context of WGD that we can set up the analytical units that enable our assessment of the presence of neighbourhood selection in Sections 7.3.1, 7.3.2 and 7.3.3 above, though these effects would exist of course whatever the source of duplicate gene pairs, and indeed in any process of genome shrinkage through gene loss.

7.4 Toward efficient, accurate guided genome halving

Algorithms for guided genome halving (GGH), or reconstruction of the pre-doubling genome with the help of an outgroup, were first used for the ancestral doubled genome of the maize (*Zea mays*) genome, with the rice (*Oryza sativa*) and sorghum (*Sorghum bicolor*) genomes as outgroups [238]. We generated all the 1.5×10^6 solutions to the genome halving problem for the maize genome, and then identified the subset, containing only a handful of relatively similar solutions that have a minimum rearrangement distance with the rice (or sorghum) genome.

This approach was feasible with the small number (34) of doubled blocks identified in maize that were also present in one copy in each outgroup, but in a subsequent analysis [174], when we attempted to reconstruct the ancient doubled yeast genome from which *Saccharomyces cerevisiae* is descended, guided simultaneously by both of the undoubled outgroup genomes *Ashbya gossypii* and *Kluyveromyces waltii*, the number of doubled genes we could use as evidence was an order of magnitude greater than the number of blocks in the cereals data, and the number of solutions to the

halving problem astronomical. It was no longer feasible to exhaustively search the halving solutions to find those that are closest to the outgroups. Instead we took a random sample of several thousand solutions in the hope that the best one might be optimal, or close to it. It was not clear, however, how large the sample should be, or how to validate the results, since the local optima found in that study remained fairly far apart, as measured by genomic rearrangement distance.

In our current use of GGH, on yeast [241] and on the flowering plants studied in the present article, we seek to replace the brute force approach of generating all (or a random sample of) halving solutions first, i.e., before taking into consideration the outgroup genome. Instead, we inject all pertinent information derivable from the outgroup into the halving algorithm, influencing hitherto arbitrary choices in that algorithm so that the halving solution is guided towards the outgroup.

7.4.1 Definitions: Genomes, rearrangement operations and genomic distance

A genome G is represented by a set of strings (called *chromosomes*) of form $\{g_{11}\cdots g_{1n_1}, ..., g_{\chi 1}\cdots g_{\chi n_\chi}\}$, where $n = n_1 + \cdots + n_\chi$ and $\{|g_{..}|\} = \{1, \cdots, n\}$; i.e., each integer $i \in \{1, \cdots, n\}$ appears exactly once in the genome and may have either positive or negative polarity. The biologically-motivated operations of reversal or inversion, reciprocal translocation, chromosome fission or fusion, and transposition, can all be represented by an operation (called double-cut and join, or DCJ) of cutting the genome twice, each time between two elements on one of the chromosomes and rejoining the four resulting cut ends differently [17,231]. Whether the two cuts are on the same chromosome or not, and how the endpoints are rejoined, determine which rearrangement operation pertains.

The genome rearrangement distance $d(G, H)$ is defined to be the minimum number of DCJ operations required to convert one of the genomes, G, into the other, H.

Rearrangement algorithms [13, 210, 231] can be formulated in terms of the bi-coloured "breakpoint graph", where each end (either 5′ or 3′) of a gene in genome G is represented by a vertex joined by a black edge to the vertex for adjoining end of the adjacent gene, and these same ends, represented by the same $2n$ vertices in the graph, are joined by gray edges determined by the adjacencies in genome H. In addition, each vertex representing a first or last term of some chromosome in G or in H is connected by an edge of the appropriate colour to an individual "cap" vertex, and there are specific rules for adding caps to the genome with fewer chromosomes and for joining the caps among themselves. if G has χ chromosomes and H has no more than χ, there are $2n + 4\chi$ vertices in all. The breakpoint graphs necessarily consist of disjoint alternating colour cycles, and it can be shown that, in the DCJ formulation, $d(G, H) = n + \chi - c$, where c is the number of cycles in the breakpoint graph. Calculating the distance can be done in time linear in n.

7.4.2 Genome halving

Let T be a genome consisting of ψ chromosomes and $2n$ genes $a_1^{(1)} \cdots, a_n^{(1)}; a_1^{(2)}, \cdots, a_n^{(2)}$, dispersed in any order on the chromosomes. For each i, we call $a_i^{(1)}$ and $a_i^{(2)}$ "duplicates", but there is no particular property distinguishing all elements of the set of $a_i^{(1)}$ in common from all those in the set of $a_i^{(2)}$. A potential "doubled ancestor" of T is written $A' \oplus A''$, and consists of 2χ chromosomes, where some half (χ) of the chromosomes, symbolized by the A', contains exactly one of $a_i^{(1)}$ or $a_i^{(2)}$ for each $i = 1, \cdots, n$. The remaining χ chromosomes, symbolized by the A'', are each identical to one in

the first half, in that where $a_i^{(1)}$ appears on a chromosome in the A', $a_i^{(2)}$ appears on the corresponding chromosome in A'', and where $a_i^{(2)}$ appears in A', $a_i^{(1)}$ appears in A''. We define A to be either of the two halves of $A' \oplus A''$, where the superscript (1) or (2) is suppressed from each $a_i^{(1)}$ or $a_i^{(2)}$. *The genome halving problem for T is to find an A for which some $d(A' \oplus A'', T)$ is minimal.*

In the rearrangement distance algorithm, construction of the breakpoint graph is an easy step. The genome halving algorithms [51] also make use of the breakpoint graph, but the problem here is the more difficult one of building the breakpoint graph where one of the genomes (the doubled ancestor $A' \oplus A''$) is unknown. This is done by segregating the vertices of the graph in a natural way into subsets, such that all the vertices of each cycles must fall within a single subset, and then constructing these cycles in an optimal way within each subset so that the black edges correspond to the structure of the known genome T and the gray edges define the adjacencies of $A' \oplus A''$.

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

Following this, for each pair of neighbouring genes, say (a_t, a_h) and (b_h, b_t), the two adjacent vertices a_h and b_h are linked by a black edge, denoted $\{a_h, b_h\}$ in the notation of Ref. [17]. For a vertex at the end of a chromosome, say b_t, it generates a virtual edge of form $\{b_t, \text{end}\}$. Note that the use of "end" instead of "cap" reflects a somewhat different bookkeeping for the beginnings and ends of chromosome in the halving algorithm compared to the distance algorithm in Section 7.4.1.

The edges thus constructed are then partitioned into *natural graphs* according to the following principle: If an edge $\{x, y\}$ belongs to a natural graph, then so does some edge of form $\{\bar{x}, z\}$ and some edge of form $\{\bar{y}, w\}$. If a natural graph has an even

number of edges, it can be shown that in all optimal ancestral 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.

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. Thus the supernatural graph may be completed one at a time.

An important detail in this construction is that before a gray edge is added during the completion of a supernatural graph, it must be checked to see that it would not inadvertently result in a circular chromosome. Key to the linear worst-case complexity of the halving algorithm is that this check may be made in constant time.

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.

7.4.3 Genome halving with outgroups

Let T be a genome consisting of ψ chromosomes and $2n$ genes $a_1^{(1)} \cdots, a_n^{(1)}; a_1^{(2)}, \cdots, a_n^{(2)}$, dispersed in any order on the chromosomes, where for each i, genes $a_i^{(1)}$ and $a_i^{(2)}$ are duplicates. Any genome R is a reference or outgroup genome for T if it contains the n genes $a_1, \cdots, a_n$.

Let R be a reference genome for T. The GGH problem with one outgroup is to find a potential ancestral genome A such that some $d(R, A)+d(A' \oplus A'', T)$ is minimal. In practice, A is either one of the solutions to the unconstrained halving problem, or it is close to such a solution [175], so little is lost in restricting our search to the set of solutions of the genome halving problem for T.

One strategy, suitable for small data sets, as in Ref. [238], is to generate the entire set $\mathbf{S}$ of genome halving solutions of T, then to evaluate each $A \in \mathbf{S}$ to find the one that minimizes $d(R, A)$.

When $\mathbf{S}$ is so large that it is not feasible to generate all of $\mathbf{S}$ in order to find the best A, we may resort to sampling $\mathbf{S}$, as in Ref. [174]. In defining the gray edges in the supernatural graphs of Section 7.4.2, we generally have several choices at some of the steps. By randomizing this choice, we are effectively choosing a random sample of $X \in \mathbf{S}$.

7.5 The GGH algorithm

The key idea in our improvement over brute force algorithms is to incorporate information from R during the halving process. It is important to take advantage of the common structure in T and R as early as possible, before it can be destroyed in the course of construction. To this end, we drop the practice of completing all the gray edges in one supernatural graph before starting another. We simply look for elements of common structure and add gray edges accordingly, always making sure that no circular chromosomes are inadvertently created.

Missing homologs The halving algorithm requires full gene sets at several steps in reconstructing the ancestor, so we algorithmically restore the missing homologs to the

most appropriate positions in T and R at the outset. The criterion for restoring a gene to a position in a genome is the net decrease in the number of disrupted adjacencies in the three-way comparison of the augmented genomes versus the situation before the gene was restored. Note that the fictional genes thus included do not count in the main GGH algorithm when it comes to choosing among steps of equal priority.

Paths We define a path to be any connected fragment of a breakpoint graph, namely any connected fragment of a cycle. We represent each path by an unordered pair $(u, v) = (v, u)$ consisting of its current endpoints, though we keep track of all its vertices and edges. Initially, each black edge in T is a path, and each black edge in R is a path.

Pathgroups A pathgroup Γ is an ordered triple of paths, two in T and one in R, where one endpoint of one of the paths in T is the duplicate of one endpoint of the other path in T and both are orthologous to one of the endpoints of the path in R. The other endpoints may be duplicates or orthologs to each other, or not.

7.5.1 The algorithms

In adding pairs of gray edges to connect duplicate pairs of terms in the breakpoint graph of T versus $A' \oplus A''$, (which is being constructed), our approach is basically greedy, but with a sophisticated look-ahead. We can distinguish five different levels of desirability, or priority, among potential gray edges, i.e., potential adjacencies in the ancestor.

Recall that in constructing the ancestor A to be close to the outgroup R, such that $A' \oplus A''$ is simultaneously close to T, we must create as many cycles as possible in the breakpoint graphs between A and R and in the breakpoint graph of $A' \oplus A''$

versus T.

1. Adding two gray edges would create two cycles in the breakpoint graph defined by T and $A' \oplus A''$, by closing two paths. When this possibility exists, it must be realized, since it is an obligatory choice in any genome halving algorithm. It may or may not also create cycles in the breakpoint graph comparison of X with the outgroup, but this does not affect its priority.

2. Adding two gray edges would create two cycles, one for T and one for the outgroup.

3. Adding two gray edges would create a cycle in the T versus $A' \oplus A''$ comparison, but none for the outgroup. It would, however, create a higher priority pathgroup.

4. Adding two gray edges would create a cycle in the T versus $A' \oplus A''$ comparison, but none for the outgroup, nor would it create any higher priority pathgroup.

5. Each remaining path terminates in duplicate terms, which cannot be connected to form a cycle, since in $A' \oplus A''$ these must be on different (and identical) chromosomes. In supernatural graphs containing such paths, there is always another path and adding two gray edges between the endpoints of the two paths can create a cycle.

In not completing each supernatural graph before moving on to another, we lose the advantage in Ref. [51] of a constant time check against creating circular chromosomes. The worst case becomes a linear time check. This is a small liability, because the worst case scenario is seldom realized, the check almost always requiring only one or two steps.

7.6 GGH results and discussion

Our data consisted of 6144 gene sets, of which only 2104 were full sets. There were only 836 defective sets by virtue of a missing ortholog in V, while 3204 genes lacked one paralog in T.

Table 7.1 shows the results of the analysis on the full gene sets only, on combinations of the full sets with one kind of defective sets, and all three sets. For each case we study not only the reconstructed ancestor but also a "projected" version where genes from the defective sets are simply erased, in order to assess the changes in gene order due to the defective gene sets. Whereas the distance between each T and its reconstructed ancestor A is given by GGH, the distance between projected ancestor and T required a heuristic, explained in detail in Refs. [242] and [208], for attributing each paralog in T to one of the two copies of the ancestral genome. Note that we choose only one optimal ancestor A for each analysis; this does not affect d for the comparison between $A \oplus A$ and *Populus*, but it may have a very small effect on b, and r and on all three quantities for the comparison between A and *Vitis*.

Figure 7.5 depicts the result of analyzing all the 6144 gene sets with GGH, although the 836 genes with no grape ortholog are not visible. This is just one of a many equally parsimonious solutions of the GGH problem, differing largely in the how they concatenate chromosomal segments where there are two different possibilities suggested by the *Populus* genome and a third by the *Vitis* genome. The reconstruction is given as an example, and the details, including the telomeric positions determining the size of the chromosomes, are not definitive.

Despite this ambiguity in the reconstruction, the numerical results on d, b and r are quite robust, and can be used for the comparison of genomes and for evaluating methods.

Table 7.1: Comparisons of the reconstructed immediate pre-doubling ancestor A with the *Vitis* genome and of the immediate doubled ancestor $A \oplus A$ with *Populus*. PPV: full gene sets, PP: defective, missing grape ortholog, PV: defective, missing one poplar paralog. Projected: genes not in PPV ancestor deleted from solution A, d: genomic distance, b:,number of breakpoints, $r = 2d/b$: the re-use statistic.

	genes	$d(A, Vitis)$			$d(A \oplus A, Populus)$		
data sets	in A	d	b	r	d	b	r
PPV	2104	638	751	1.70	454	690	1.32
PPV,PP	2940	649	757	1.71	737	1090	1.35
projected	2104	649	757	1.71	581	823	1.41
PPV,PV	5308	1180	1331	1.77	1083	1457	1.49
projected	2104	663	758	1.75	670	833	1.61
PPV,PP, PV	6144	1208	1363	1.77	1337	1812	1.48
projected	2104	664	757	1.75	750	926	1.62
without singletons							
PPV	2020	560	661	1.69	346	541	1.28
PPV,PP	2729	594	690	1.72	453	714	1.27
projected	2006	571	664	1.72	416	628	1.32
PPV,PV	4203	573	686	1.67	751	1031	1.46
projected	1955	489	580	1.69	490	644	1.52
PPV,PP, PV	4710	675	797	1.69	856	1211	1.41
projected	1986	528	622	1.70	558	744	1.50

The large number of singleton genes disrupting otherwise homogeneous synteny blocks suggests that "noise" due to uncertainties inherent in homology identification and especially orthology identification may be artifactually inflating genomic distance d and the number of breakpoints b. Since the rigorous noise elimination techniques of Refs. [239] and [39], which are the gene-order equivalent of synteny block construction methods for genome sequences, have not yet been extended in the context of genome doubling, we simply identified singletons as gene sets lacking two real (i.e., not inferred from **insertMH**) common adjacencies out of six possible in the original genomes, and ran all the analyses again without these genes.

In each case, we counted the breakpoints and calculated the appropriate genomic distance d, i.e., from the doubled ancestor to *Populus* and from the undoubled version of the same ancestor to *Vitis*.

Algorithm GGH:
Guided Genome Halving with Full and Defective Gene Sets

Input. Two genomes: duplication descendant T', outgroup genome R', where each gene is has three
homologs (full set) or two homologs (defective set), in the patterns TTR, TT or TR.
Output. Augmented genomes T, and R, where all gene sets are full, and
Genome A, a halving solution of T, minimizing $d(A' \oplus A'', T) + d(A, R)$.
insertMH
Initialize paths (black edges) in T and R.
Construct supernatural graphs.
Construct two pathgroups for each gene g in R, one based on g_t, the other on g_h.
If number of chromosomes in T is odd,
add pathgroup with two paths of form (end, end).
While there remains at least one pathgroup
For each pathgroup $((x, y), (\bar{x}, z), (x, m))$
classify it by case and priority, and find a pathgroup Γ that has the highest priority. To choose among
Priority 2 pathgroups, find one that maximizes the number of "real" black edges, i.e., edges in T' and R',
not just edges created by **insertMH**. Similarly for Priority 3 pathgroups.
Case 1: $\bar{x} \neq y$, and adding xy and $\bar{x}\bar{y}$ would not create a circular chromosome.
Priority 1: $z = \bar{y}$.
Priority 2: $y = m$.
Priority 3: adding xy and $\bar{x}\bar{y}$ would create a pathgroup with priority 2.
Priority 4: None of 1, 2 or 3.
Case 2: $\bar{x} \neq y$, and adding $x\bar{z}$ and $\bar{x}z$ would not create a circular chromosome.
Priority 2: $z = m$.
Priority 3: adding $x\bar{z}$ and $\bar{x}z$ would create a pathgroup with priority 2.
Priority 4: Neither of 2 or 3.
Case 3:$\bar{x} = y$.
Priority 5:
If Γ is Case 1, **addGrayEdge**$(xy, \bar{x}\bar{y})$.
If Γ is Case 2, **addGrayEdge**$(x\bar{z}, \bar{x}z)$.
If Γ is Case 3, find some
$W = ((w, \bar{w}), (\bar{w}, w), (w, s))$ in the same supernatural graph and
and **addGrayEdge**$(xw, \bar{x}\bar{w})$.

Algorithm: addGrayEdge$(rt, \bar{r}\bar{t})$

Add gray edges $rt, \bar{r}\bar{t}$ to partially completed genome $X" \oplus X''$.
Add gray edge rt to partially completed genome X.
Update paths in pathgroups that are affected by the new gray edges.
Remove pathgroups that start with r and t.

Algorithm: insertMH:
Insert Missing Homologs in Chromosomes

Input. Two genomes: duplication descendant T', outgroup R', where each gene is has two or three
homologs, in the patterns TTR, TT, TR.
Output. Augmented genomes T and R containing exactly three homologs for each gene, in the pattern TTR,
maximizing the number of common edges of form $\{a_1, b_1\}$, $\{a_2, b_2\}$ in T and $\{a, b\}$ in R.
(Or $\{a_1, b_2\}$, $\{a_2, b_1\}$ in T and $\{a, b\}$ in R.)
While there are genes that have only two copies, **count edgeDiff** for each such, which simultaneously finds
the BestPosition.
Insert the gene with the minimum edgeDiff value into the BestPosition of this gene.

Algorithm: count edgeDiff

If a gene g just has one copy (g_1) in T' and one copy (g) in R', then we must insert another copy (g_2) into T'.
If a gene g just has two copies (g_1, g_2) in T', then we must insert g into R'.

(The details are omitted here. This is essentially a greedy heuristic to add adjacencies reflecting, as if possible, adjacencies already existing in R' and T'.)

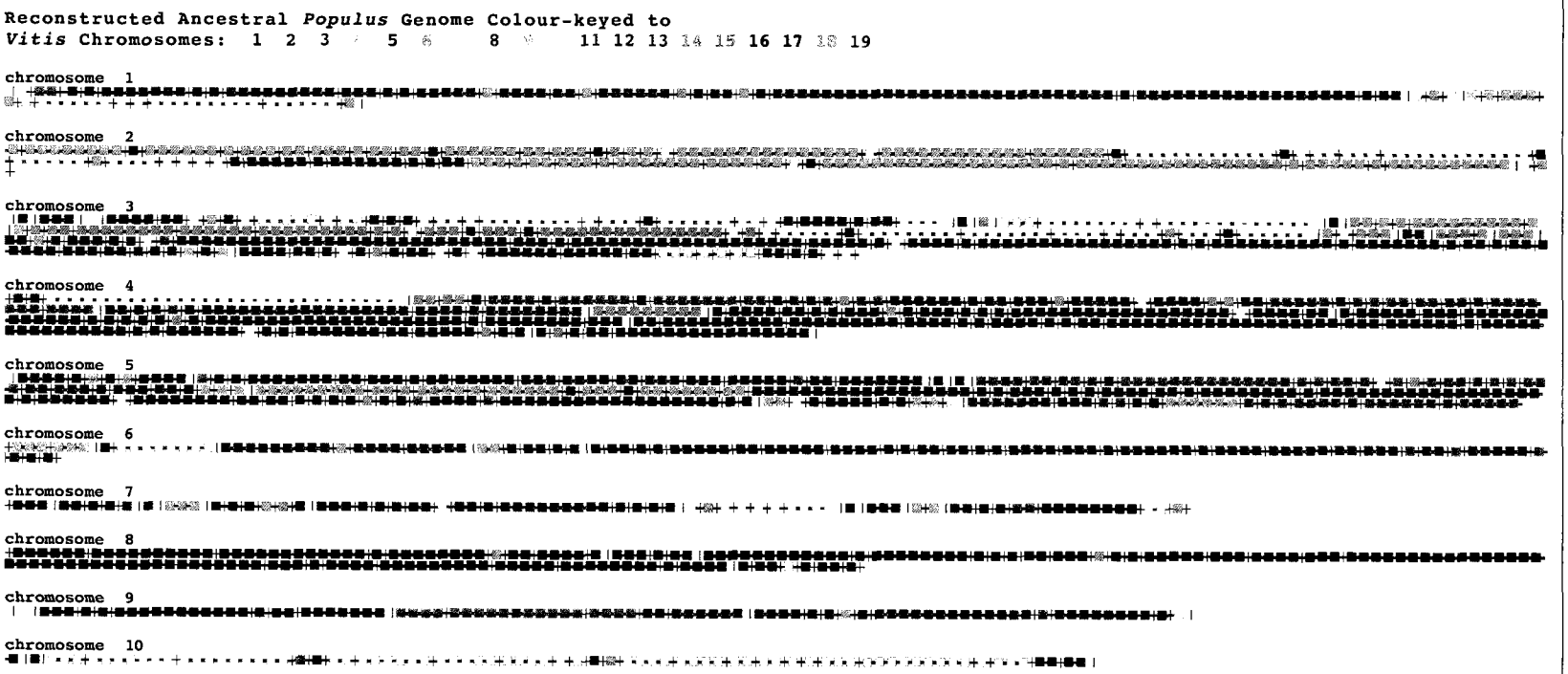

Figure 7.5: Ten chromosomes (wrapped) of ancestral poplar genome reconstructed by GGH algorithm from 6144 full and defective gene sets. Only genes with grape orthologs are indicated. Adjacencies present three times, i.e., twice in poplar and once in grape are indicated by $\equiv$, those present twice by $=$ and those present once by $-$. Intrachromosomal breakpoints within segments indicated by $|$.

This enabled us to calculate the "breakpoint re-use" statistic $r = 2d/b$, which is a measure of how much signal about conserved order (among segments, not within segments) remains in the comparison of two genomes after a period of evolutionary rearrangements. When $r = 1$, we can have high confidence in the rearrangement distance and history. When r approaches two, the segment order in the two genomes being compared are essentially random with respect to each other, i.e., calculating r for random genomes gives a value approaching 2[1]. In Table 7.1, we see both from changes in d and changes in r that

- most of the signal contained in the order among conserved chromosomal segments has been lost between the ancestor and *Vitis*, but is retained to a great degree between the ancestor and *Populus*, probably reflecting the difference in divergence time but also possible biases towards T in the GGH algorithm [242],
- the addition of the defective PV gene sets degrades the analysis, more than the addition of PP sets, though this may due to the four times greater number of gene sets in the former,
- the elimination of singletons improves all the analyses, but where PV is present, this comes about largely by discarding most of the sets, which turn out to be singletons.

The analysis with 6144 gene sets required almost 48 hours on a MacBook, but this was anomalously large, since those with 4000 or 5000 required less than 5 hours and those with 2000 about 1 hour. Much of the running time is due to the check on the number of real edges in a pathgroup to choose among Priority 2 or among Priority 3 options. This could

[1]If breakpoints are frequently re-used during evolution, then r will also be close to 2; unfortunately there is no internal way of testing the breakpoint re-use hypothesis against the null hypothesis of complete loss of signal about segment order [173].

be reduced by optimizing data structures in our software.

7.7 Conclusions

We have formalized a new way of assessing neighbourhood selection constraints on duplicate gene loss, taking into account the particular genomic structure of descendants of WGD. Dividing possible selective effects into those affecting which gene pairs will lose one member, and those affecting which member of the pair is lost, we found no effect of the first type, but a clear effect of the second.

With the application of our GGH method to the more than 6000 gene sets, we have shown that any realistic case of genome doubling should be amenable, even if all the gene paralogs remain in the sequenced descendant.

The inclusion of defective PV gene sets would appear to add little more than noise to the analysis, but the PP sets would seem to add significant information, especially to the ancestor-*Populus* comparison.

The elimination of singletons proves to be a meaningful way of drastically decreasing the number of segments (as measured by b) and the genomic distance to credible levels, though this still does not result in a detectible signal in the ancestor-*Vitis* comparison.

The recently sequenced *Carica papaya* genome [135], which is phylogenetically more closely related to *Populus*, but like *Vitis* diverged before the *Populus* doubling event, can also play the outgroup role in our analysis [175], but accuracy is diminished since genome assembly has not been completed. In general, as sequences become more polished and further complete homology sets can be more accurately detected, our methods should become

more accurate.

Acknowledgments

Research supported in part by grants from the Natural Sciences and Engineering Research Council of Canada (NSERC). DS holds the Canada Research Chair in Mathematical Genomics.

Chapter 8

Towards improved reconstruction of ancestral gene order in angiosperm phylogeny

Sankoff D, Zheng C, Wall PK, dePamphilis C, Leebens-Mack J, Albert V. Journal of Computational Biology. 2009. Submitted on December 19, 2008. [Conference version: Sankoff D, Zheng C, Wall PK, dePamphilis C, Leebens-Mack J, Albert V.2008. Internal validation of ancestral gene order reconstruction in angiosperm phylogeny. Proceedings of RECOMB-CG 2008: 252–264.]

I was responsible for designing the algorithm and getting all of the results in this work. I also collaborated fully in writing and preparing the manuscript, especially technical write-ups and graphics. P.K. Wall, a member of C. dePamphilis' research team, provided the *Populus*, *Caria* and *Vitis* data organized into homology sets. Some of the work and

manuscript revision was done in consultation with V.A. Albert, and J. Leebens-Mack also provided helpful information.

Abstract

Whole genome doubling (WGD), a frequent occurrence during the evolution of the angiopsperms, complicates ancestral gene order reconstruction due to the multiplicity of solutions to the genome halving process. Using the genome of a related species (the outgroup) to guide the halving of a WGD descendant attenuates this problem. We investigate a battery of techniques for further improvement, including an unbiased version of the guided genome halving algorithm, reference to two related genomes instead of only one to guide the reconstruction, use of draft genome sequences in contig form only, incorporation of incomplete sets of homology correspondences among the genomes and addition of large numbers of "singleton" correspondences. We make use of genomic distance, breakpoint reuse rate, dispersion of sets of alternate solutions and other means to evaluate these techniques, while reconstructing the pre-WGD ancestor of *Populus trichocarpa* as well as an early rosid ancestor.

8.1 Introduction

The reconstruction of the gene order in ancestral genomes requires that we make a number of choices, among the data on which to base the reconstruction, in the algorithm to use and in how to evaluate the result. In this paper we illustrate an approach to making these choices in the reconstruction of the ancestor of the poplar *Populus trichocarpa* genome. This species that has undergone whole genome duplication followed by extensive chromosomal rearrangement, and is one of four angiosperm genomes, along with those of *Carica papaya* (papaya), *Vitis vinifera* (grapevine) and *Arabidopsis thaliana*, whose sequences have been published to date, shown in Figure 8.1.

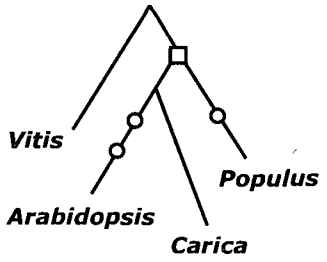

Figure 8.1: Phylogenetic relationships among angiosperms with sequenced genomes. The circles indicate likely whole genome doubling events. The circle in the *Populus* lineage, representing the locus of the WGD event at the origin of the willow-poplar family, and the square, representing the ancestor of the rosid dicotyledons, indicate the target ancestors we reconstruct in this paper.

We have been developing methods to incorporate descendants of whole genome doubling into phylogenies of species that have been unaffected by the doubling event. The basic tool in analyzing descendants of whole genome doubling is the halving algorithm [51]. To overcome the propensity of the genome halving procedure to produce numerous, widely disparate solutions, we "guide" the execution of this procedure with information from genomes

from related species [174,238,241,242,244], which we call outgroups. This, *ipso facto*, integrates the whole genome doubling descendant into the phylogeny of the related species.

Issues pertaining to data include

Homology sets. Can we use defective sets of homologs, i.e., which have only one copy in the duplicated genome or are missing the ortholog completely in the guide genome?

Singletons. Should we purge singletons from the data, i.e., sets of homologous markers who have no homologous adjacent markers in common in the either the duplicated genome or the outgroup?

Contigs. Can we use guide genomes that are not fully assembled, but are available only as sets of hundreds or thousands of contigs?

Another choice to be made during reconstruction has to do with the guided halving algorithm itself. The original genome halving problem, with no reference to outgroup genomes, can be solved in time linear in the number of markers [51]. We can introduce information from an outgroup in order to guide this solution, without compromising the optimality of the result and without serious increase in computing time [241, 244]. We call this *constrained* guided halving. The true, *unconstrained*, guided halving problem, however, where the solution ancestor need not be a solution of the original halving problem, is likely to be NP-hard [208]. In the heuristics necessary for these two approaches, there is a trade-off between the speed of constrained halving versus the (theoretically) better solution obtainable by unconstrained halving.

Once we make our choices of data and algorithm, we may ask how to evaluate the

results. As with most evolutionary reconstructions, this evaluation is necessarily completely internal, since there is no outside reference to check against, except simulations. There are many indices for evaluating a reconstruction.

Distance. Most important, there is the objective function; here our genomic distance definition attempts to recover the most economical explanation of the observed data, namely the minimum number of rearrangement events (reversals, reciprocal translocations, chromosome fusions/fissions, transpositions) required.

Reuse rate. Each rearrangement operation can create at most two breakpoints in the gene-by-gene alignment of two genome and its ancestor. When fewer than two are created, one or two pre-existing breakpoint(s) must be "re-used". Conversely, when rearranged genomes are optimally reconstructed, some breakpoints may be reused. In fact breakpoint re-use is inferred far more frequently in reconstruction than it actually occurs in genome generation, and is actually a measure of the loss of evolutionary signal inherent in the gene order.

Dispersion. The motivation for guided halving is to resolve the ambiguities inherent in the large number of optimal halving solutions. One way to quantify the remaining non-uniqueness is to calculate the distances among a sample of solutions. Thus we can compare the average distance between the alternate solutions in one method to the average in another, to see which is the less dispersed, or more compact. And we can compare these "within-group" distances to "between-group" distances, to assess statistically how much our methodological choices affect the results.

In this paper we will refer repeatedly to a main tabulation of results, Table 8.1, in which

we discover the unexpected rapid evolution of the *Carica* gene order in comparison with that of *Vitis*. In Section 8.2, we report on the origin and processing of our gene-order data and the construction of the full and defective homology sets. In Section 8.3 we take up the discussion of our measures for assessing the quality of reconstructions. Then, in Section 8.4, we discuss the halving problems, and sketch a new algorithm for unconstrained guided halving. In Section 8.5 we evaluate the utility of singletons and of defective homology sets. Then, in Section 8.6 we assess the two guided halving algorithms on real and simulated data. Section 8.7 proposes a way to use unassembled genome sequence in contig form, as a input to the reconstruction algorithm, an approach that could potentially have wide use in gene order phylogeny. In Section 8.8 we demonstrate the phylogenetic validity of reconstructing the *Populus* ancestor using either *Vitis* or *Carica*, or both, as outgroups. Note that we have not included *Arabidopsis* in our analyses; as will be explained in Section 8.9, this was dictated by a paucity of data in the appropriate configurations.

8.2 The Populus, Vitis and Carica data

Annotations for the *Populus*, *Vitis* and *Carica* genomes were obtained from databases maintained by the U.S. Department of Energy's Joint Genome Institute [213], the French National Sequencing Center, Genoscope [81], and the University of Hawaii [135], respectively. An all-by-all BLASTP search was run on a data set including all *Populus* and *Vitis* protein coding genes, and orthoMCL [113] was used to construct 2104 full and 4040 defective gene sets, in the first case, denoted PPV, containing two poplar paralogs (genome P) and one grape ortholog (genome V), and in the second case, denoted PV or PP, missing a copy from

either P or V. This was repeated with *Populus* and *Carica*, genomes P and C, respectively, to obtain 2590 full (PPC) and 4632 defective (PC or PP) sets. The location on chromosomes (or contigs in the case of *Carica*) and orientation of these paralogs and orthologs was used to construct our database of gene orders for these genomes. Contigs containing only a single gene were discarded from the *Carica* data.

8.3 Evaluation of solutions

Developing methods for historical inference about genomes is an uncertain enterprise, since there is usually no way of checking the results against historical truth, the fossil record being extremely fragmentary, vast evolutionary time scales generally precluding laboratory experimentation and simulation being extremely dependent on simplifying assumptions.

Nevertheless there are meaningful evaluation criteria. Parsimonious explanations are to be preferred to uneconomical ones or, if there is an accepted probability model, most likely explanations are better than unlikely ones. Low variance estimates are better than high variance ones. And methods that allow internal tests of significance, e.g. the bootstrap, are desirable. Here we will discuss the three sorts of evaluation we use in this study.

8.3.1 Genome distance and breakpoint graph

The distance measures we use are based on parsimony. As such they are likely to produce underestimates of the number of rearrangements historically intervening between two genomes, especially if this number is large. Nevertheless, lacking a credible probabilistic model for rearrangement processes, we can rely on the current measures, as long as we do

data sets	genes in A, with singletons	$d(A, Vitis)$ d	b	r	$d(A \oplus A, Populus)$ d	b	r	total d
Solutions constrained to also be solutions of genome halving								
PPV	2104	638	751	1.70	454	690	1.32	1092
PPV,PP	2940	649	757	1.71	737	1090	1.35	1386
PPV,PV	5308	1180	1331	1.77	1083	1457	1.49	2263
PPV,PP, PV	6144	1208	1363	1.77	1337	1812	1.48	2545
Solutions unconstrained								
PPV	2104	593	734	1.62	512	733	1.40	1105
PPV, PP	2940	616	752	1.64	778	1119	1.39	1394
PPV,PV	5308	1121	1307	1.72	1147	1486	1.54	2268
PPV,PP,PV	6144	1129	1328	1.70	1437	1871	1.54	2566
data sets	genes in A, with singletons	$d(A, Carica)$ d	b	r	$d(A \oplus A, Populus)$ d	b	r	total d
Solutions constrained to also be solutions of genome halving								
PPC	2590	896	1075	1.67	565	823	1.37	1461
PPC, PP	3478	905	1085	1.67	884	1282	1.38	1789
PPC,PC	6334	1892	2224	1.70	1262	1700	1.48	3154
PPC,PP,PC	7222	1925	2241	1.72	1541	2065	1.49	3466
Solutions unconstrained								
PPC	2590	864	1043	1.66	628	870	1.44	1492
PPC, PP	3478	873	1039	1.68	951	1318	1.44	1824
PPC,PC	6334	1859	2172	1.71	1321	1742	1.52	3180
PPC,PP,PC	7222	1877	2211	1.70	1617	2126	1.52	3494
data sets	genes in A, without singletons	$d(A, Vitis)$ d	b	r	$d(A \oplus A, Populus)$ d	b	r	total d
Solutions constrained to also be solutions of genome halving								
PPV	2020	560	661	1.69	346	541	1.28	906
PPV,PP	2729	594	690	1.72	453	714	1.27	1047
PPV,PV	4203	573	686	1.67	751	1031	1.46	1324
PPV,PP, PV	4710	675	797	1.69	856	1211	1.41	1531
Solutions unconstrained								
PPV	2020	545	652	1.67	375	564	1.33	920
PPV, PP	2729	567	681	1.67	493	745	1.32	1060
PPV,PV	4203	544	674	1.61	782	1034	1.51	1326
PPV,PP,PV	4710	631	785	1.61	916	1250	1.47	1547
data sets	genes in A, without singletons	$d(A, Carica)$ d	b	r	$d(A \oplus A, Populus)$ d	b	r	total d
Solutions constrained to also be solutions of genome halving								
PPC	2464	772	934	1.65	412	607	1.36	1184
PPC, PP	3226	812	981	1.66	536	809	1.33	1348
PPC,PC	4651	779	926	1.68	774	1050	1.47	1554
PPC,PP,PC	5234	898	1088	1.65	892	1253	1.42	1790
Solutions unconstrained								
PPC	2464	758	921	1.65	454	639	1.42	1212
PPC, PP	3226	796	967	1.65	584	839	1.39	1380
PPC,PC	4651	764	911	1.68	804	1090	1.48	1568
PPC,PP,PC	5234	861	1058	1.63	952	1303	1.46	1813

Table 8.1: Guided halving solutions with and without singletons, constrained and unconstrained heuristics, *Vitis* or *Carica* as outgroup, and all combinations of full and defective homolgy sets. A: pre-doubling ancestor of *Populus*, $A \oplus A$: doubled ancestor, PPV, PPC: full gene sets, PP: defective, missing grape or papaya ortholog, PV,PC: defective, missing one poplar paralog. d: genomic distance, b, number of breakpoints, $r = 2d/b$: the reuse statistic.

not forget the inherent bias towards "shorter" solutions.

Genome comparison algorithms generally involve manipulations of the bicoloured breakpoint graph [13, 210] of two genomes, called the black and the gray genomes, on the same set of n genes, where two vertices are defined representing the two ends of each gene, and an edge of one colour joins two vertices if the corresponding gene ends are adjacent in the appropriate genome. Omitting the details pertaining to the genes at the ends of chromosomes, the genomic distance d, i.e., the minimum number of rearrangements necessary to transform one genome into the other, satisfies $d = n - c$, where c is the number of alternating colour cycles making up the breakpoint graph [231].

It is well-known [131, 184] that in practice genomic distance depends strongly on the degree of resolution of the genomic data. The smaller the threshold for conserved segment size and the greater the number of segments, the greater the distance. This is true for d and n as well. We will have to take account of the dependence of d on n when we investigate the effects of singletons, types of homology class, algorithm version and outgroup in Section 8.5.

8.3.2 Breakpoint reuse

If d is the number of rearrangements and b the number of breakpoints, the reuse [154] variable $r = 2d/b$ can take on values in $1 \leq r \leq 2$. Completely randomized genomes will have r close to 2, so that if an empirical comparison has $r \sim 2$, we cannot ascribe much significance to the details of the reconstruction [173]. This is particularly likely to occur for genomes that are only very distantly related. In fact, studies of mammalian genomes [184] have shown a very close correlation between r and d. This does not indicate an actual

tendency towards breakpoint re-use throughout a phylogenetic domain, *since then r would be elevated even for closely related genomes*, but rather a loss of gene-order signal due to inadequate modeling of evolutionary processes and/or the reconstruction of homologous gene orders [173].

8.3.3 Dispersion

As we will see in Section 8.4, algorithms for reconstructing ancestral genomes generally allow two or more choices at many stages. It suffices to make this choice randomly to generate a sample of alternative solutions. The distances between these solutions are suggestive of the reliability of the method. A method that produces solutions within a few rearrangements of each other is preferable to one that generates a set of very heterogeneous solutions, as long as this improvement does not come with the cost of an increased bias.

8.4 Guided halving

The genome halving problem [51] 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 labeling as "1" or "2" of the two genes in a pair of copies, so that all n copies labeled "1" are in one half of $A \oplus A$ and all those labeled "2" are in the other half. The genome A represents the ancestral genome at the moment immediately preceding the WGD event giving rise to

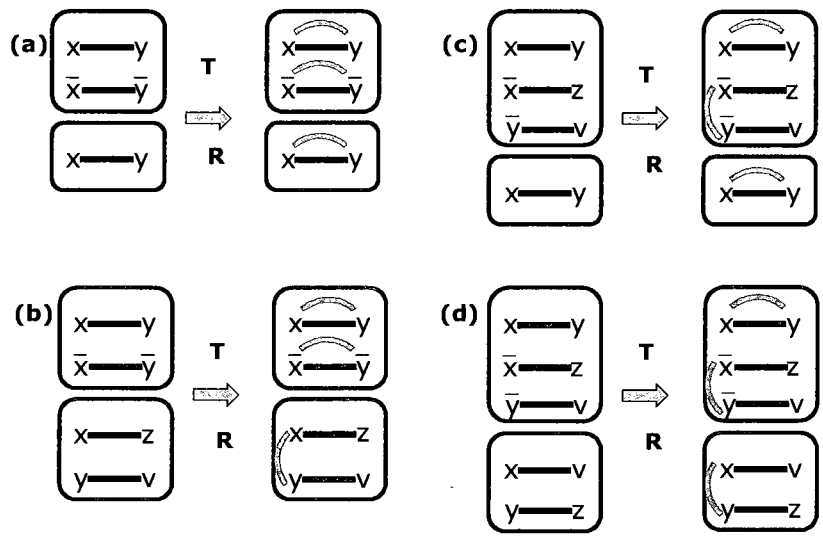

Figure 8.2: Choice of gray edge to add at each stage of the reconstruction of A and $A \oplus A$. Each black edge in the diagram represents either an adjacency in T or R or an alternating colour path with a black edge at each end point. If vertex w is copy "1" in T then $\bar{w}$ is copy "2", and vice versa. (a) Configuration requiring the creation of three cycles, two in the breakpoint graph of T and $A \oplus A$, and one in the breakpoint graph of A and R. (b) Configuration requiring the creation of two cycles in the breakpoint graph of T and $A \oplus A$, necessary for $A \oplus A$ to be a solution of the genome halving problem. (c) Alternative configuration if solution of guided halving $A \oplus A$ is not also required to be a solution of the halving problem. (d) Look-ahead when there are no configurations (a), (b) or (c). Here the addition of three gray edges creates a configuration (c).

$A \oplus A$.

The guided genome halving problem [238] 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.

In previous studies [174, 238, 241], we found that 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. 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 halv-

ing 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. Without sketching out the details of the lengthy algorithm, the addition of gray edges representing genome A to the breakpoint graph, as in Figure 8.2, must favour configuration (b) over (c), even though there are as many cycles created by (c) as by (b). This is a consequence of the original halving theory in Ref. [51]. Otherwise $A \oplus A$ may not be a halving solution. This, however, may bias the reconstruction of A towards T and away from R. For the present work, we implemented a new version of the algorithm, as sketched in Section 8.4.1, treating configurations (b) and (c) equally in constructing A. The choice among two or more configurations of form (b) or (c) is based on a look-ahead calculation of what effect this choice will have on the remaining inventory of configurations of form (b) and (c). The new algorithm requires much more computation, but its objective function is better justified.

8.4.1 The new algorithm

First we define paths, which represent intermediate stages in the construction of the breakpoint graph comparing T and $A \oplus A$ and the breakpoint graph comparing A and R. Then we define pathgroups, which focus on the three current paths leading from three "homologous" vertices in the graph, namely two copies in T and one in R. Note that each vertex represents one of the two ends of a gene.

Paths We define a path to be any connected fragment of a breakpoint graph, namely any connected fragment of a cycle. We represent each path by an unordered pair $(u, v) = (v, u)$ consisting of its current endpoints, though we keep track of all its vertices and edges. Initially, each black edge in T is a path, and each black edge in R is a path.

Pathgroups A pathgroup, as in Figure 8.2, is an ordered triple of paths, two in the partially constructed breakpoint graph involving T and $A \oplus A$ and one in the partially constructed breakpoint graph involving R and A, where one endpoint of one of the paths in T is the duplicate of one endpoint of the other path in T and both are orthologous to one of the endpoints of the path in R. The other endpoints may be duplicates or orthologs to each other, or not.

In adding pairs of gray edges to connect duplicate pairs of terms in the breakpoint graph of T versus $A \oplus A$, (which is being constructed), our approach is basically greedy, but with a careful look-ahead. We can distinguish four different levels of desirability, or priority, among potential gray edges, i.e., potential adjacencies in the ancestor.

Recall that in constructing the ancestor A to be close to the outgroup R, such that $A \oplus A$ is simultaneously close to T, we must create as many cycles as possible in the breakpoint graphs between A and R and in the breakpoint graph of $A \oplus A$ versus T. At each step we add three gray edges.

- Priority 1. Adding the three gray edges would create two cycles in the breakpoint graph defined by T and $A \oplus A$, by closing two paths, and one cycle in the breakpoint graph comparison of A with the outgroup, as in Figure 8.2a.

- Priority 2. Adding three gray edges would create two cycles, one for T and one for the outgroup, or two for T and none for the outgroup, as in Figure 8.2b and c.

- Priority 3. Adding the gray edges would create only one cycle, either in the T versus $A \oplus A$ comparison, or in the R versus A comparison. In addition, it would create a higher priority pathgroup, as in as in Figure 8.2d.

- Priority 4. Adding the gray edges would create only one cycle, but would not create any higher priority pathgroup.

The algorithm simply completes the steps suggested by the highest priority pathgroup currently available, choosing among equal priority pathgroups according to a look-ahead to the configuration of priorities resulting from competing moves.

At each step, we must verify that a circular chromosome is not created, otherwise the move is blocked. As with Ref. [51] this check requires a constant time. The algorithm terminates when no more pathgroups can be completed. Any remaining pathgroups define additional chromosomes in the ancestor A.

8.5 On the utility of singletons and defective homology sets

From the last column of Table 8.1, it is clear that d varies widely as a function of the four factors, inclusion/exclusion of singletons, inclusion/exclusion of defective homology sets, outgroup species and heuristic. But it is also clear that d depends on n, in the first numerical column in the table (cf. [131, 184]). Thus we must control for the dependence of d on n in

teasing out the relative contribution of each of these factors. In Figure 8.3 we group the 32 points in the plot of d versus n, taken from the 32 rows in Table 8.1, according to choice of outgroup, inclusion or not of singletons, combination of homology classes and algorithm version. We will return to the almost imperceptible differences between the constrained and unconstrained algorithms in Section 8.6, and to the choice of outgroup in Section 8.8, but we can observe here that the inclusion singletons has a dramatic effect on the rate of increase of d on n. Though this effect can theoretically be generated by rearrangements, in practice it is better considered as noise in the analysis [39, 239]. This is confirmed by the greater values of r, indicating degradation of evolutionary signal, almost everywhere in the upper half of Table 8.1 compared to the lower half.

We also note that the increase in d caused by adding defective homology sets to the analysis is really due to the disproportionate numbers of singletons in these sets. The trend lines for the four different combinations of homology sets are parallel and steeply sloped. This slope is largely due to the presence of singletons in the data for the two highest point on each line and also to the higher rate of evolution of *Carica* for the highest and third highest points. The actual effect of homology class can be traced by comparing the lowest points on the four lines, the second lowest points, and so on. This shows a relatively gradual increase.

8.6 Comparison of the heuristics

In Table 8.1, the constrained guided halving algorithm always does better than the unconstrained guided halving heuristic, as measured by the total distance in the last column. At

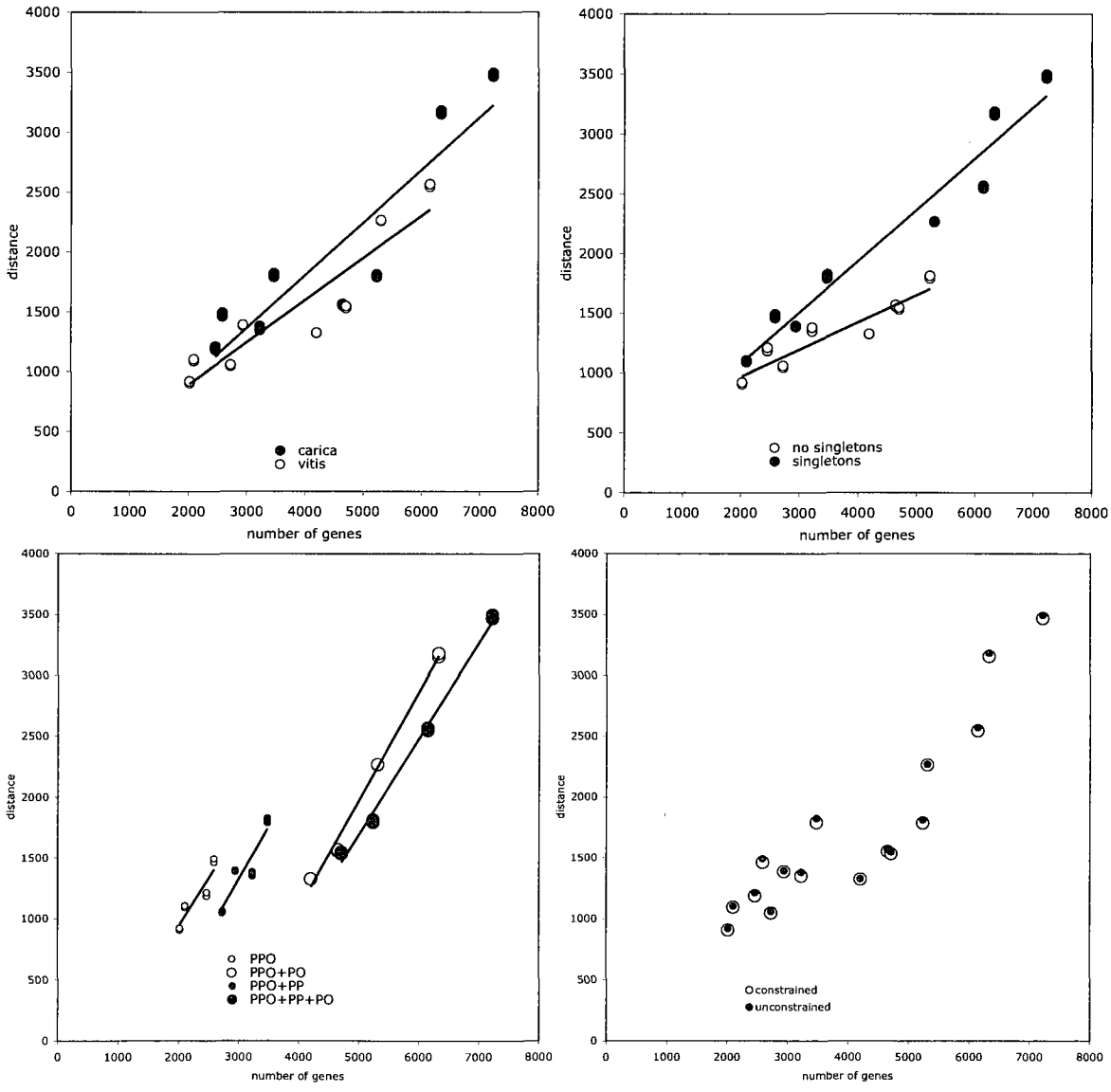

Figure 8.3: Effect of controlling for the number of genes. Upper left: *Carica* evolving faster than *Vitis*. Upper right: rapid increase in distance due to singletons. Lower left: Effect of homology classes. Lower right: almost imperceptible effect of algorithm version.

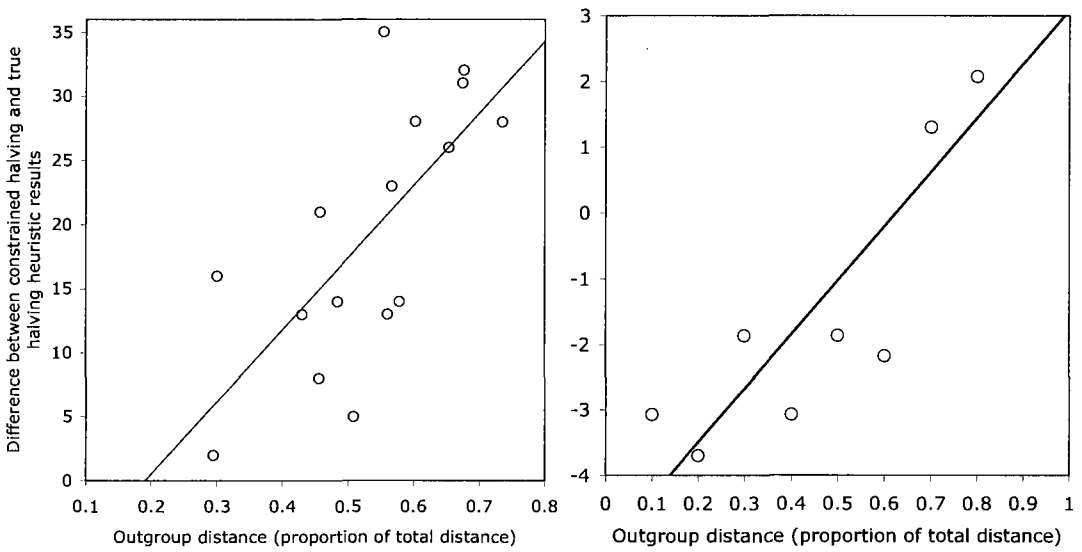

Figure 8.4: Performance of the constrained and unconstrained heuristics as a function of the real (left) or simulated (right) distance of the outgroup from A. Note that despite the similarity of the two curves, the simulated results indicate that the new (unconstrained) algorithm is better when the outgroup proportion of total distance is no larger than than 0.6, while with the real data this is only predicted to happen when that proportion is below 0.2.

the same time, the unconstrained heuristic had a clear effect in reducing the bias towards *Populus*, in each case decreasing the distance to the outgroup, compared to the constrained heuristic. This decrease was accompanied by a small decrease in r for the outgroup analysis.

In fact the decrease in the bias was far greater than the increase in total cost, meaning that if bias reduction is important, then this heuristic is worthwhile, despite its inability to find a minimizing ancestor and its lengthy execution time.

To further investigate the behaviour of the new algorithm, we simulated evolution by M inversions and translocations (in a 10:1 proportion) from a genome A to produce an outgroup genome R and $1000 - M$ rearrangements from a WGD genome $A \oplus A$ to produce a descendant genome T. We then applied the constrained and the new algorithms, showing

that the new one was superior when $M \leq 600$, but not for $M \geq 700$, as seen in Figure 8.4 (right).

Considering the 16 comparisons in the real data between the constrained and the new algorithm, the change in the total distance also shows a distinct correlation ($\rho^2 = 0.5$) with the distance from the outgroup and A. We point this out even though the constrained algorithm, as we have seen, seems superior when the distance between R and A is more than 20 % of the total distance. This is plotted in Figure 8.4 (left).

The difference between the simulations, where the new method is generally superior, and the real data, where the new method would seem to be superior only when the outgroup is very close to the ancestor, must be ascribed in large part to some way the model used for the simulations does not correspond to how the real data was generated. The "failure" of the new algorithm to do better with the real data cannot be ascribed to its inability to find good local optima, since it succeeds with simulated data. One clue is the relatively high reuse rate in the comparison between the outgroup and A, compared with that between *Populus* and $A \oplus A$.

8.7 Rearrangements of partially assembled genomes

Our analyses involving *Carica* have incorporated an important correction. The genomic distance between *Carica* and A counts many chromosome fusion events that reduce the number of "chromosomes" in *Carica* from 223 to the 19. These are not a measure of the true rearrangement distance, but only of the current state of the *Carica* data. Since these may be considered to take place as a first step in the rearrangement scenario [231], we may

	genes	$d(A, Carica)$			correction						
data sets	in A	d	b	r	c	d^*	a	ct	ca	b^*	r^*
PPC	2464	986	1090	1.81	223	772	76	7	1371	934	1.65
PPC, PP	3226	1027	1132	1.81	224	812	74	6	2091	981	1.66
PPC,PC	4651	1084	1177	1.84	314	779	123	9	3470	926	1.68
PPC,PP,PC	5234	1214	1318	1.84	325	898	112	12	3910	1088	1.65

Table 8.2: Correction for contig data. A: pre-doubling ancestor of *Populus*, $A \oplus A$: doubled ancestor, PPC: full gene sets, PP: defective, missing papaya ortholog, PC: defective, missing one poplar paralog. d: genomic distance, b: number of breakpoints, $r = 2d/b$: the reuse statistic, c: number of contigs, d^*: distance corrected for excess of contigs over true number of chromosomes $= d - c + 9$, a: number of "obvious fusions", ca: number of common adjacencies, ct: number of common telomeres, b^*: corrected number of breakpoints = number of genes $-ca - ct - 2a$, r^*: corrected reuse statistic $= 2d^*/b^*$. Data without singletons. Solutions obtained by constrained algorithm.

simply subtract their number from d to estimate the true distance. At the same time, many of the breakpoints between A and *Carica* are removed by these same fusions, so these should be removed from the count of b as well. The calculations in Table 8.2 illustrate how the $d(A, Carica)$ results in the bottom quarter of Table 8.1 were obtained.

Figure 8.5 (left) shows experimental results on how the increasing fragmentation of a genome into contigs, using a random fragmentation of *Vitis* grenome, decreases the estimated distance between *Vitis* and A. This is understandable, since the freedom of the contigs to fuse in any order without this counting as a rearrangment step, inevitably will reduce the distance by chance alone. But the linearity of the result suggests that this decrease is quite predictable, and that the estimates of the distance between *Carica* and A are actually underestimates by about 10 %.

Figure 8.5 (right) shows that creating contigs by randomly breaking the *Vitis* genome does not create excessive variability among the solutions, only the same as the dispersion of

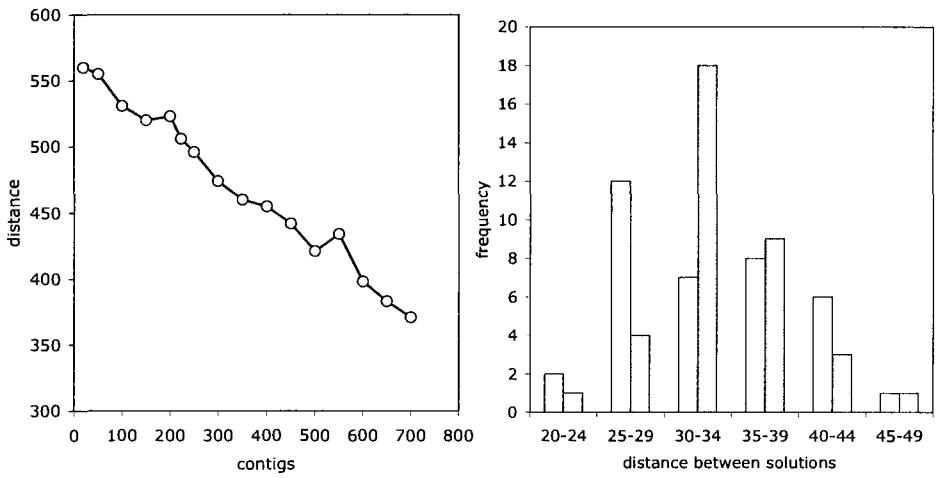

Figure 8.5: Left: Effect of increasing fragmentation of *Vitis* into "contigs" on the distance between the reconstructed A and *Vitis*. Right: Distributions of distances among solutions for A based on *Vitis* data (white bars) and among solutions for *Vitis* fragmented into contigs in different random ways (gray bars).

alternate solutions for the original *Vitis* data, a few percentage points of the distance itself.

8.8 A comparison of the outgroups

Perhaps the most surprising result in Table 8.1 is that the *Vitis* gene order is decidedly closer to *Populus* and its ancestor A than *Carica* is. Both the Tree of Life and the NCBI Taxonomy Browser currently exclude the Viticae family from the rosids, though some older taxonomies do not make this distinction.

Before interpreting this result, we mention two sources of error in the comparison of *Vitis* and *Carica*. The first is that the *Carica* distances are based on a larger gene set; without singletons and defective homology sets PPC is 22 % larger than PPV. As a rule of thumb, we can expect distances to be approximately proportional to the number of genes. However, as we have seen in Figure 8.3, *Carica* evolves faster even if we control for gene

number.

The other source of error is due to the contig data, and this results in an *underestimate* of the *Carica*-ancestor distance. From Figure 8.5, we can estimate that the *Carica* distances are underestimated by about 10 % because of the 223 contigs in the *Carica* data. Thus the discrepancy between the two outgroups is actually larger than it appears to be.

We may conclude that this difference is genuine and substantial. Then assuming that *Populus* and *Carica* have a closer phylogenetic relationship, or even a sister relationship, our results can only be explained by a faster rate of gene order evolution in *Carica* than in *Vitis*.

8.8.1 Dispersion

As described in Section 8.3.3, we generated 100 different solutions with the constrained halving and unguided halving algorithms using each outgroup, and 54 for unconstrained halving with *Vitis* as the outgroup and 15 with *Carica*. The genomic distances were normalized by number of genes in common in the two genomes being compared before input to the analysis. This number was 2464 genes for *Carica* comparisons, 2020 for *Vitis* comparisons and 1514 for *Carica*-*Vitis* comparison.

For each of the six outgroup/method combinations we calculated the average normalized distance between all of its solutions to each of the other combinations, leading to the 6×6 matrix in Table 8.3. We input this into a two-dimensional principal coordinates analysis in the R package, producing the pattern of six black dots in Figure 8.6. It can be seen that the first dimension of the figure represents the right-to-left movement from unguided halving towards increasing influence of the outgroup. The second dimension distinguishes

	Carica	*Vitis*	ConC	ConV	UncC	UncV	UgC	UgPV
Carica	0.0	458.4	400.2	432.6	394.5	432.6	448.9	463.7
Vitis		0.0	289.3	278.7	305.2	270.3	313.7	329.7
ConC			14.2	52.2	40.6	57.5	68.6	81.9
ConV				16.8	73.3	35.1	76.6	68.3
UnvC					16.6	69.4	94.6	107.0
UncV						20.3	89.2	87.1
UgC							45.0	63.4
UgV								47.0

Table 8.3: Matrix of average distances between analyses, normalized $\times 1000$. Diagonal contains average within-group distances (not input into principal components analysis). *Carica* and *Vitis* data used in Figure 8.7 but not in Figure 8.6. Con = constrained, Unc = unconstrained, Ug = unguided, C = *Carica*, V = *Vitis*.

the *Carica* and *Vitis* analyses.

We then calculated the average distance between all the alternate solutions *within* each outgroup/method combination and divided this by the original input distances to the vertical and horizontal neighbours of the corresponding point on Figure 8.6; these two factors were multiplied by the corresponding distances on the principal coordinates graph in order to obtain the axes of ellipses. The ellipses, shaded in the figure, represent the degree of dispersion of the solutions around each of the six points. In the case of the constrained solution based on *Vitis*, a quadrilateral shape was employed because of the asymmetry of the horizontal comparisons involving the unconstrained guided halving and the unguided halving solutions.

Figure 8.7 situates the ancestral reconstructions in a principal coordinates analysis including the *Carica* and *Vitis* genomes. The *Populus* genome is added after the analysis on the basis of halving distances; it was not included in the principal coordinates analysis because of orthology assignment inconsistencies arising in the calculations of the distances

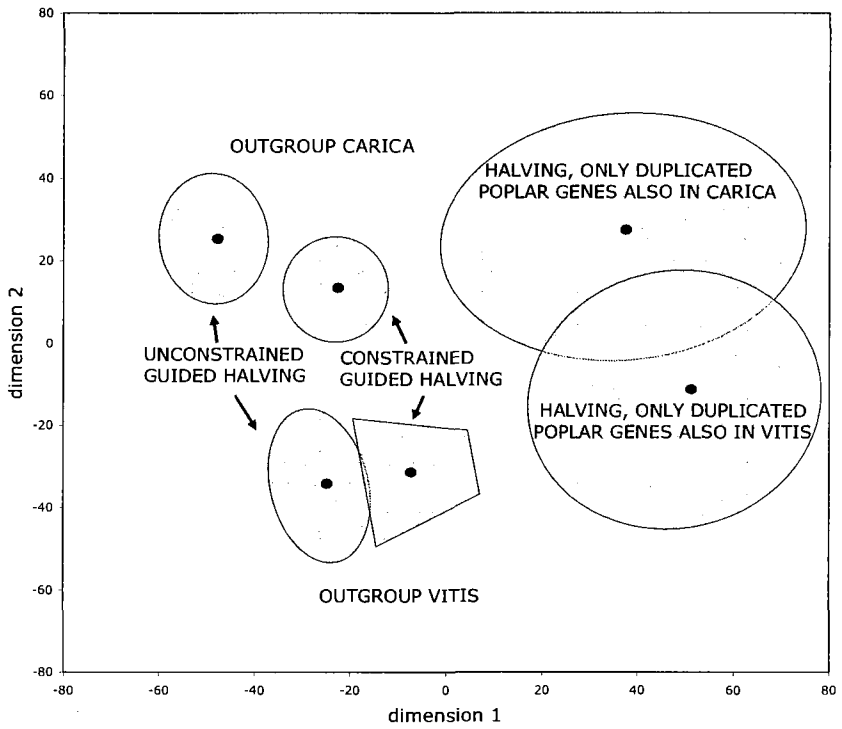

Figure 8.6: Principal coordinates analysis of average distances between reconstructed ancestral genomes. Black points represent "average" genomes. Shaded areas around each point represent the dispersion of alternate solutions of the same halving problem.

between several unduplicated genomes and the descendant of a WGD.

This figure shows that the ancestral reconstructions all occupy a relatively small area of solution space. It also represents the movement already studied in Figure 8.6 from unguided to constrained to unconstrained analysis in the direction of the outgroup.

8.8.2 Using both outgroups

There are 1734 complete homologous gene sets including two *Populus* copies and one copy in each of *Carica* and *Vitis*. (Some of these, constituting 1-gene contigs in *Carica*, were not used for the analyses in Table 8.1. Here we have 332 *Carica* contigs, instead of the 223 in the

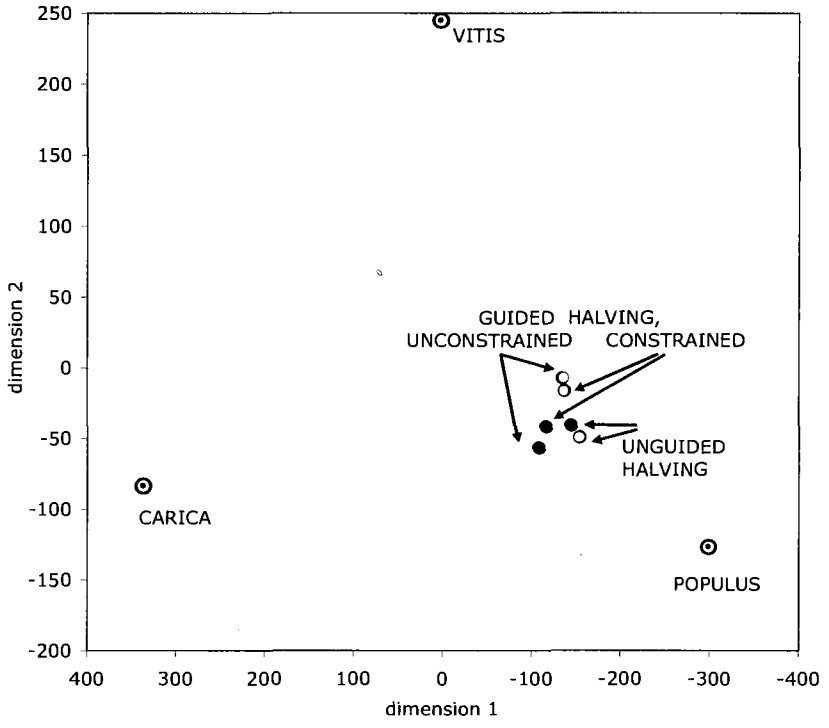

Figure 8.7: Principal coordinates analysis of distances between reconstructed and present-day genomes.

previous analysis.) In the same way as the unconstrained algorithm in Section 8.4 is based on a modification of the guided halving algorithm for one outgroup in reference [241], we could define an unconstrained version of the two-outgroup guided halving algorithm implemented in that earlier work. For convenience, however, we use the constrained version of two-outgroup guided halving from reference [241] to find the ancestor (small circle) genome in Figure 8.8(a) as a first step, then compute the "median" genome based on this ancestor, *Carica* and *Vitis*. The median problem here is to find the genome, the sum of whose distances from ancestor A, *Carica* and *Vitis* is minimal. This problem is NP-hard [208] and solving it is barely feasible with the 1734 genes in our data, requiring some 300 hours of MacBook computing time.

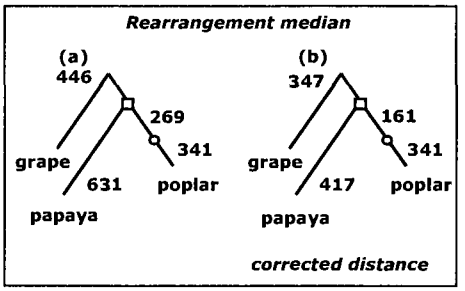

Figure 8.8: Branch lengths in angiosperm phylogeny, using two estimates of the median, and applying the contig correction.

This initial result unfortunately inherits the same defect as the *Carica* data, i.e., it is composed of contigs rather than true chromosomes. In this case, the median genome contains 118 "contig-chromosomes". We correct the distances to the median by subtracting the difference in the number of chromosomes/contigs between the three genomes and the median. This corresponds to disregarding the fusions counted in the original distances that are essentially carrying out an optimal assembly, modeling an analytical process, not a biological one. This produces the corrected values in Figure 8.8(b).

Let us compare the distance from *Vitis* and from *Carica* to ancestor A, passing through the median, in Figure 8.8 (508 and 578, respectively), with the minimum distances[1] in Table 8.1, and proportionately adjusted for the reduced number of genes ($560 \times \frac{1734}{2020} = 481$ and $772 \times \frac{1734}{2464} = 543$, respectively. Passing through the median modestly augments (by 27 and by 35, respectively) both trajectories. But using the median diminishes the total cost of the phylogeny, i.e., in comparison with a phylogeny where there is no common evolutionary divergence of the outgroups from *Populus* from $481 + 543 = 1024$ to $347 + 417 + 161 = 925$.

Figure 8.8(b) confirms that the papya genome has evolved more rapidly than the

[1]Constrained analyses, no singleton or defective homology sets.

grapevine one.

8.8.3 Molecular evolutionary correlates of rearrangement rates

With obvious sources of error in our papaya/poplar and three-way comparisons (such as the incomplete assembly of the papaya genome and potentially error-prone ortholog/paralog determination) insufficient explanations for papaya's enhanced rearrangement rate relative to *Vitis* or the diploid poplar ancestor, A, we have sought a biological interpretation.

Papaya, grapevine, and poplar all share the ancient γ WGD. "Paleologous" (paralogous) gene pairs identified as γ descendants, as mined from the three genomes, show different rates (K_s) of synonymous substitutional change [206]. Median K_s for *Vitis* γ pairs (1.22) is substantially lower than that for poplar (1.54) or papaya (1.76).

Synonymous substitutional rates can be interpreted as placeholders for divergence times [41], but they have also been correlated with different life strategies in plants, e.g., the woody perrenial versus the annual habit, and as such, generation time [59]. Recent evidence from large-scale phylogenetic studies incorporating many taxa and many genes has backed the latter inference [185].

The generation times of papaya, poplar, and grapevine show a pattern entirely (negatively) consistent with median K_s values for γ paralogs. Papaya can reproduce in 9-15 months [135], poplar in 4-6 years [213], and grape (sexually) in ca. 80 years [11]. In turn, these generation times and K_s values correlate well with the genomic rearrangement rates calculated here (Table 8.1; Figure 8.8). As such, we hypothesize a common cause argument, short generation time, to explain the aberrant-seeming rearangement history for papaya relative to its phylogenetic relationships, which would otherwise have suggested this taxon to

be closer than grapevine to the diploid poplar ancestor, A. Future median genome/GGH analyses incorporating weedy species such as *Arabidopsis* and *Mimulus* might help bolster or refute this hypothesis depending on their rearrangement rates relative to other plant genomes.

8.9 Conclusions

The main contributions of this paper are:

- the discovery of the rapid rate of gene order evolution in *Carica* compared to *Vitis*,
- a systematic way of controlling for the dependence of rearrangement distance on the number of genes,
- a way of visualizing the reduction of dispersion of the solutions to a problem when comparing methods to solve it,
- a way to use incompletely assembled contigs in genome rearrangement studies,
- a new unbiased algorithm for guided genome halving, and
- the systematic use of reuse rates to show that the inclusion of singletons are not helpful in ancestral genome reconstruction.

In this work, we have not considered the *Arabidopsis* genome. The main reason is not any algorithmic issue, but the paucity of full homology sets containing four *Arabidopsis* copies as well as copies from one or more outgroups.

Acknowledgments

Research supported in part by a Discovery grant to DS and a doctoral fellowship to CZ, from the Natural Sciences and Engineering Research Council of Canada (NSERC). DS holds the Canada Research Chair in Mathematical Genomics.

Chapter 9

Multichromosomal median and halving problems under different genomic distances

Eric Tannier, Chunfang Zheng, David Sankoff. 2008. Submitted to BMC Bioinformatics on December 15, 2008. [Conference version: Workshop on Algorithms in Bioinformatics WABI 2008, Lecture Notes in Bioinformatics 5251:1–12, Springer.]

Most of the work in this paper was done by Dr. Eric Tannier, though in continuous consultation with our team. We also provided the section on the breakpoint distance. The inspiration for this work was the solution of the complexity problem in my earlier paper [241]. I verified all the mathematics in the paper, correcting and changing much of it in the process. I collaborated in preparing the manuscript and some of the graphics.

Abstract

Background: Genome median and genome halving are combinatorial optimization problems that aim at reconstructing ancestral genomes as well as the evolutionary events leading from the ancestor to extant species. Exploring complexity issues is a first step towards devising efficient algorithms. The complexity of the median problem for unichromosomal genomes (permutations) has been settled for both the breakpoint distance and the reversal distance. Although the multichromosomal case has often been assumed to be a simple generalization of the unichromosomal case, it is also a relaxation so that complexity in this context does not follow from existing results, and is open for all distances.

Results: We settle here the complexity of several genome median and halving problems, including a surprising polynomial result for the breakpoint median and guided halving problems in genomes with circular and linear chromosomes, showing that the multichromosomal problem is actually easier than the unichromosomal problem. Still other variants of these problems are NP-complete, including the DCJ double distance problem, previously mentioned as an open question. We list the remaining open problems.

Conclusions: This theoretical study clears up a wide swathe of the algorithmical study of genome rearrangements with multiple multichromosomal genomes.

9.1 Background

The gene order or syntenic arrangement of ancestral genomes may be reconstructed based on comparative evidence from present-day genomes — the phylogenetic approach — or on internal evidence in the case of genomes descended from an ancestral polyploidisation event, or from a combination of the two. The computational problem at the heart of phylogenetic analysis is the *median problem*, while internal reconstruction inspires the *halving problem*, and the combined approach gives rise to *guided halving*. How these problems are formulated depends (1) on the karyotypic framework: the number of chromosomes in a genome and whether they are constrained to be linear, or if circular chromosomes are also permitted, and (2) on the objective function used to evaluate possible solutions. This function is based on some notion of genomic distance, either the number of adjacent elements on a chromosome in one genome that are disrupted in another — the breakpoint distance — or the number of evolutionary operations necessary to transform one genome to another.

While the karyotypes allowed in an ancestor vary only according to the dimensions of single versus multiple chromosome, and linear versus circular versus mixed, the genomic distances of interest have proliferated according to the kinds of evolutionary operations considered, from the classic, relatively constrained, reversals/translocations distance to the more inclusive *Double Cut-and-Join* (DCJ) measure, and many others [52].

The computational complexity of some of these problems has been settled for some specific distances and karyotypic contexts, and it is sometimes taken for granted that these results carry over to other combinations of context and distance. This is not necessarily the case. In this paper, we survey the known results and unsolved cases for three distance

measures in three kinds of karyotype. We include several results presented here for the first time, as well as discussions on the definitions of the distances. The results contain both new polynomial-time algorithms and NP-hardness proofs. This paper is the full version of an extended abstract that has appeared in [208], which announced the results without giving all the proofs. In particular, a full discussion on the breakpoint distance definition, as well as the proofs of Theorem 3, Theorem 5, and Theorem 7 are added here, which makes this version a complete and definitive one.

9.1.1 Genomes, breakpoints and rearrangements

Multichromosomal genomes

We follow the general formulation of a genome in [17]. A *gene* A is an oriented sequence of DNA, identified by its *tail* A^t and its *head* A^h. Tails and heads are the *extremities* of the genes. An *adjacency* is an unordered pair of gene extremities. A *genome* Π is a set of adjacencies on a set of genes. Each adjacency in a genome means that two gene extremities are consecutive on the DNA molecule. In a genome, each gene extremity is adjacent to zero or one other extremity. An extremity x that is not adjacent to any other extremity is called a *telomere*, and can be written as an adjacency $x\circ$ with a null symbol $\circ$. The adjacency $x\circ$ is called a *telomeric adjacency*. For a genome Π on a set of genes $\mathcal{G}$, consider the graph G_Π whose vertices are all the extremities of the genes, and the edges include all the non telomeric adjacencies in Π as well as an edge joining the head and the tail of each gene. This graph is a set of disjoint paths and cycles. Every connected component is called a *chromosome* of Π. A chromosome is *linear* if it is a path, and *circular* if it is a cycle. A

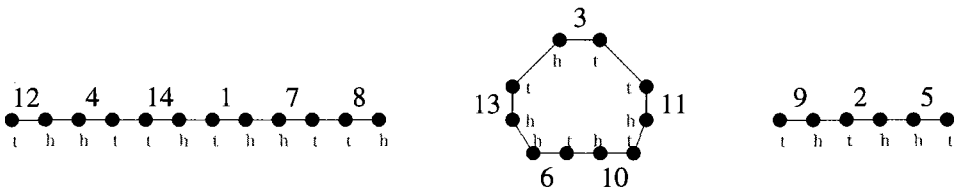

Figure 9.1: The graph G_Π of a genome Π. Π is a genome on the set of genes $\{1, \ldots, 14\}$, containing three chromosomes, two of them being linear and one circular. Its adjacencies are the union of
$C_1 = \{12^h 4^h, 4^t 14^t, 14^h 1^t, 1^h 7^h, 7^t 8^t\}$,
$C2 = \{3^t 11^t, 11^h 10^t, 10^h 6^t, 6^h 13^h, 13^t 3^h\}$ and $C_3 = \{9^h 2^t, 2^h 5^h\}$. It has four telomeres.

genome with only linear, or only circular, chromosomes is called a *linear* or *circular* genome, respectively. An example of a graph G_Π is given in Figure 9.1.

A Genome can also be represented as a set of strings, by writing the genes for each chromosome in the order in which they appear in the paths and cycles of the graph G_Π, with a bar over the gene if the head of the gene appears before the tail (we say it has *negative* sign), and none if the tail appears before the head (it has *positive* sign). For each linear chromosome, there are two possible equivalent strings, according to the arbitrary chosen starting point. One 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 cycle is traversed. For example, chromosome C_1 of the genome Π of Figure 9.1 may be written $(12\ \bar{4}\ 14\ 1\ \bar{7}\ 8)$ or $(\bar{8}\ 7\ \bar{1}\ \bar{14}\ 4\ \bar{12})$.

A genome with only one chromosome is called *unichromosomal.* These correspond to *signed permutations*: the two string representations are (linear or circular) signed permutations.

Genomes with duplicates

A *duplicated gene* A is a couple of homologous oriented sequences of DNA, identified by two tails $A1^t$ and $A2^t$, and two heads $A1^h$ and $A2^h$. An *all-duplicates genome* Δ is a set of adjacencies on a set of duplicated genes.

For a genome Π on a gene set $\mathcal{G}$, a *doubled genome* $\Pi \oplus \Pi$ is an all-duplicates genome on the set of duplicated genes from $\mathcal{G}$ such that if $A^x B^y$ $(x, y \in \{t, h\})$ is an (possibly telomeric) adjacency of Π (A^x or B^y may be $\circ$), either $A1^x B1^y$ and $A2^x B2^y$, or $A2^x B1^y$ and $A1^x B2^y$, are adjacencies of $\Pi \oplus \Pi$.

Note the difference between a general all-duplicates genome and the special case of a doubled genome: the former has two copies of each gene, while in the latter these copies are organised in such a way that there are two identical copies of each chromosome when we ignore the 1's and 2's in the $A1^x$'s and $A2^x$'s: it has two linear copies of each linear chromosome, and for each circular chromosome, either two circular copies or one circular chromosome containing the two successive copies[1]. Note also that for a genome Π, there is an exponential number of possible doubled genomes $\Pi \oplus \Pi$ (exactly two to the power of the number of non-telomeric adjacencies in Π). These definitions correspond to duplicated and perfectly duplicated genomes found in [136], and slightly differs from the perfectly duplicated genome definition found in [219], as discussed in [136]. An example of an all-duplicates genome and a doubled genome is shown in Figure 9.2. Doubled genomes are the immediate result of an evolutionary event called *Whole Genome Duplication* (WGD), which is known to have occurred in many evolutionary lineages, from protists [12] to yeasts, to plants, to

[1]The construct containing both copies of the doubled genome on a single circular chromosome was introduced for the sake of exploring the extension of the halving problem to the case of a single chromosome [51]. It was not motivated by biological considerations.

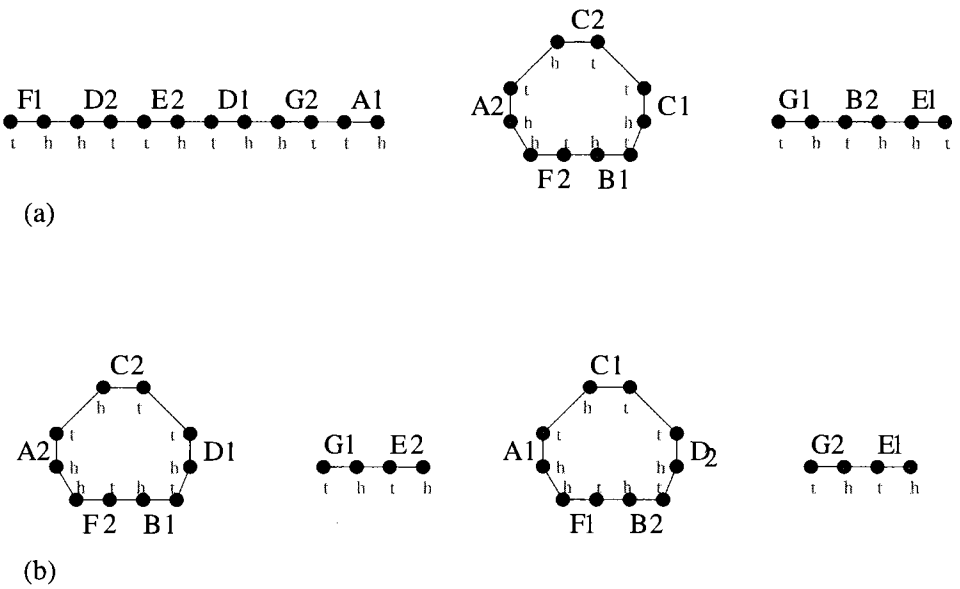

Figure 9.2: The graphs and G_Δ $G_{\Pi\oplus\Pi}$ of an all-duplicates genome Δ and a doubled genome $\Pi\oplus\Pi$ (a) Δ has three chromosomes, while (b) $\Pi\oplus\Pi$ has four, and consist of two copies of two chromosomes, that have the same set of duplicated genes $\{1,\ldots,7\}$.

fish, to amphibians and even to mammals [147]. All-duplicates genomes derive from doubled genomes through a series of rearrangement events. Typically, all-duplicates genomes pertain to extant species, while doubled genomes are ancestral configurations inferred to exist immediately after the WGD, and that are to be reconstructed.

In discussing all-duplicates genomes, we will sometimes contrast them with *ordinary genomes* which have a single copy of each gene.

The breakpoint distance

The breakpoint distance has been well-studied for permutations, i.e., unichromosomal genomes [169, 220], but only a few published discussions have focused on how it should be defined for multichromosomal genomes (see [155] for one suggestion). The distance should depend not only on common adjacencies, or rather their absence, but also on common telomeres (or

lack thereof) in two genomes. For two genomes Π and Γ on a set $\mathcal{G}$ of n genes, suppose Π has N_Π chromosomes, and Γ has N_Γ chromosomes. Let $a(\Pi, \Gamma)$ be the number of common adjacencies, $e(\Pi, \Gamma)$ be the number of common telomeres of Π and Γ. Then insofar as it should depend additively on these components, we may suppose the breakpoint distance has form

$$d_{BP}(\Pi, \Gamma) = n - a(\Pi, \Gamma)\beta - e(\Pi, \Gamma)\theta + (N_\Pi + N_\Gamma)\gamma + (|N_\Pi - N_\Gamma|)\psi,$$

where β, θ and γ are positive parameters, while ψ may have either sign. Taking $\Pi = \Gamma$ and imposing $d_{BP}(\Pi, \Pi) = 0$ yields the relations $\beta = 1$ and $1 - 2\theta + 2\gamma = 0$, so $\theta = \gamma + 1/2$, and the distance formula reduces to:

$$d_{BP}(\Pi, \Gamma) = n - a(\Pi, \Gamma) - (\gamma + 1/2)e(\Pi, \Gamma) + (N_\Pi + N_\Gamma)\gamma + (|N_\Pi - N_\Gamma|)\psi.$$

It is most plausible to count a total of 1 breakpoint for a fusion or fission of linear chromosomes[2], which implies $\gamma = \psi = 0$, so the most natural choice of *breakpoint distance* between Π and Γ is

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

This differs somewhat from the parameters chosen in [155], where the breakpoint distance is defined with $\gamma = \frac{1}{2}$ and $\psi = \frac{1}{2}$, giving rise to the disadvantage of there possibly being more breakpoints between two genomes than adjacencies in either one. For example, in comparing

[2]It might be argued that a fission or fusion should count for as many as 2 breakpoints, or anything between 1 and 2, so that alternate values of γ and ψ might be entertained, provided $\gamma \in [0, \frac{1}{2}]$, and $\psi \in [0, 1 - \gamma]$. This may have an influence on how to calculate the number of breakages within a scenario, as discussed in [18]. We do not explore this further in this paper.

$\Pi = (1\ 2\ 3\ 4\ 5)$ and Γ in which five linear chromosomes each contain one gene $i \in \{1, \ldots, 5\}$, the definition in [155] would count 9 breakpoints, which seems counterintuitive, while our definition counts 4, which seems more reasonable. Whether all the results presented in this paper also hold for the definition in [155] is open.

The definition of the breakpoint distance is easily transposable to the comparison of two all-duplicates genomes. For one all-duplicates genome Δ and one ordinary genome Π, the *breakpoint distance* between Π and Δ is the minimum breakpoint distance between Δ and a doubled genome $\Pi \oplus \Pi$, that is,

$$d_{BP}(\Pi, \Delta) = \min_{\Pi \oplus \Pi} d_{BP}(\Pi \oplus \Pi, \Delta).$$

The Double Cut-and-Join distance

Given a genome Π, a double-cut-and-join (DCJ) is an operation ρ acting on two adjacencies pq and rs (possibly some of p, q, r, s are $\circ$ symbols, so that telomeric adjacencies are considered; one adjacency can even be $\circ\circ$). The DCJ operation replaces pq and rs either by pr and qs, or ps and qr. An example of DCJ operation on the genome Π of Figure 1 is drawn in Figure 9.3.

A DCJ can reverse an interval of a genome, may cause the fission of one chromosome into two, or the fusion of two chromosomes into a one, or a reciprocal translocation: the exchange of two telomere-containing segments between two chromosomes. Two consecutive DCJ operations, excising and circularising a chromosomal segment followed by a re-linearisation of the *circular intermediate* and reintegration on the same chromosome, using two new cut-points, results in a *block interchange*: two segments of the genome appear

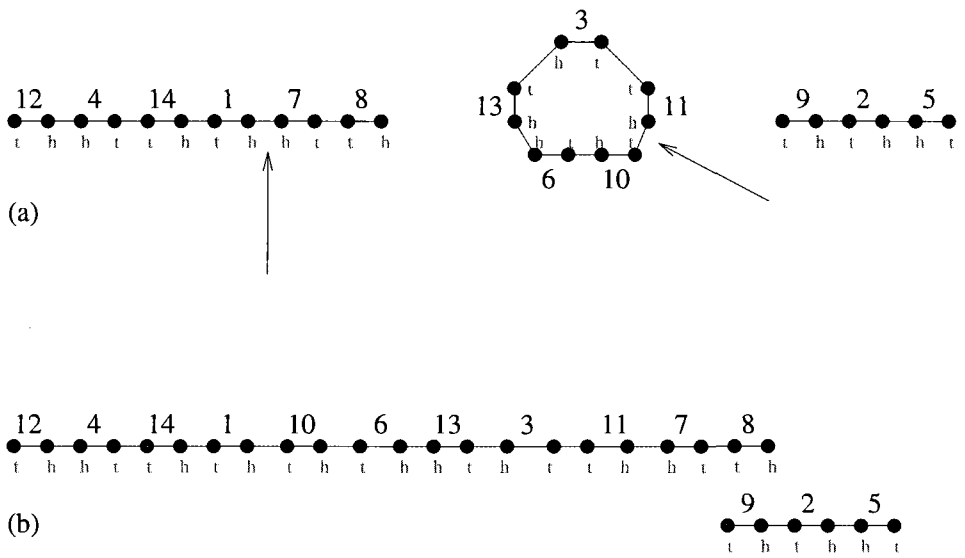

Figure 9.3: A DCJ operation on the genome Π of Figure 9.1. Adjacencies 1^h7^h and 10^t11^h of the genome represented in (a) are cut and adjacencies 1^h10^t and 11^h7^h are joined to construct the genome represented in (b). This example shows how a DCJ operation can fuse two chromosomes into one.

to simply exchange their positions. In the case these two segments are consecutive, the two DCJs result in a *transposition*, the apparent movement of a segment from one place on a chromosome to another. The DCJ operation is thus a very general framework, introduced by Yancopoulos *et al.* [231], as well as by Lin *et al.* in a special case [114], and since been adopted by Bergeron *et al.* [17, 19] and many others, sometimes under other names such as SPRING [115] or "2-break rearrangement" [7].

If Π and Γ are two genomes on a set $\mathcal{G}$ of n genes, the minimum number of DCJ operations needed to transform Π into Γ is called the *DCJ distance* and noted $d_{DCJ}(\Pi, \Gamma)$.

This DCJ distance is easily defined also for two all-duplicates genomes. For one all-duplicates genome Δ and one ordinary genome Π, the *DCJ distance* between Π and Δ is $d_{DCJ}(\Pi, \Delta) = \min_{\Pi \oplus \Pi} d_{DCJ}(\Pi \oplus \Pi, \Delta)$.

The reversal/translocation distance

The reversal/translocation distance was introduced by Hannenhalli and Pevzner [70], and is equivalent to the DCJ distance constrained to linear genomes.

If Π is a linear genome, a *linear* DCJ operation is a DCJ operation on Π that results in a linear genome. This allows reversals, chromosome fusions, fissions, and reciprocal translocations. DCJs that create circular intermediates, temporary circular chromosomes, and thereby mimic block interchanges and transpositions, are not allowed. Chromosome fusions and fissions are particular cases of translocations in this framework, justifying the appellation *RT-distance*. If Π and Γ are linear genomes, the *RT* distance between Π and Γ is the minimum number of linear DCJ operations that transform Π into Γ, and is noted $d_{RT}(\Pi, \Gamma)$.

9.1.2 Computational problems

The classical literature on genome rearrangements aims at reconstructing the evolutionary events and ancestral configurations that explain the differences between the organization of extant genomes. The focus has been on the genomic distance, median and halving problems. More recently the doubled distance and guided halving problems have also emerged as important. In each of the ensuing sections of this paper, these five problems are examined for a specific combination of distance d (breakpoint, DCJ or RT) and kind of multichromosomal karyotype (linear, circular, mixed).

1. **Distance.** Given two genomes Π, Γ, compute $d(\Pi, \Gamma)$. Once the distance is calculated, an additional problem in the cases of DCJ and RT is to reconstruct the

rearrangement scenario of length $d(\Pi, \Gamma)$, i.e. the putative events that differentiate the genomes.

2. **Double distance.** Given an all-duplicates genome Δ and an ordinary genome Π, compute $d(\Delta, \Pi)$. This computation counts the evolutionary events posterior to a WGD of the given genome Π, leading to an all-duplicates genome Δ, and locates the genes of the all-duplicates genome on chromosomes in one of the two ancestral copies of the ordinary genome. Because the assignment of labels "1" or "2" to the two identical (for our purposes) copies of a duplicated gene in Δ is arbitrary, the double distance problem is equivalent to finding such an assignment that minimises the distance between Δ and a genome $\Pi \oplus \Pi$ considered as ordinary genomes, where all the genes on any one chromosome in $\Pi \oplus \Pi$ are uniformly labeled "1" or "2" [7,242]. The double distance function is not symmetric because Δ is an all-duplicates genome and Π is an ordinary one, thus capturing the presumed asymmetric temporal and evolutionary relationship between the ancestor Π and the present-day genome Δ.

3. **Median.** Given three genomes Π_1, Π_2, Π_3, find a genome M which minimises $d(\Pi_1, M) + d(\Pi_2, M) + d(\Pi_3, M)$. The median problem estimates the common ancestor of two genomes, given a third one as an outgroup. This is meaningful even in the "unrooted" case, where it is not specified which of the three genomes is the outgroup, because of the symmetry of the sum to be minimised.

4. **Halving.** Given an all-duplicates genome Δ, find an ordinary genome Π which minimises $d(\Delta, \Pi)$, the double distance mentioned above. The goal of a halving analysis is to reconstruct the ancestor of an all-duplicates genome at the time of a

WGD event.

5. **Guided halving.** Given an all-duplicates genome Δ and an ordinary genome Π, find an ordinary genome M which minimises $d(\Delta, M) + d(M, \Pi)$. The guided halving problem is similar to the genome halving problem for Δ, but it takes into account the ordinary genome Π of an organism presumed to share a common ancestor with M, the reconstructed undoubled ancestor of Δ[3].

We will survey these five computational problems for the three distances that we have introduced, in the cases of multichromosomal genomes containing all linear chromosomes, all circular chromosomes, or permitting both. The latter are refered as *mixed genomes.*

While many problems are open for multichromosomal genomes, there is a huge amount of research on these problems for unichromosomal genomes, whether circular or linear (the two cases are often equivalent up to some transformations [52]). They are not systematically particular cases of the multichromosomal problems, as the constraint of keeping only one chromosome along a rearrangement scenario can result in more difficult problems. More precisely, unichromosomal DCJ problems reduce to RT multichromosomal ones. Indeed, the RT operations always transform a unichromosomal genome into a unichromosomal one. As this paper contains very few results on the RT distance, practically the unichromosomal cases are often independant and not generalized here. Results on unichromosomal genomes are summarised in Table 9.1, together with the results for the multichromosomal case we review or present here. A complete survey on these problems can be found in [52].

[3]A variant of the guided halving problem introduced in [238] is to find an ordinary genome M that is a solution to genome halving, that is, minimises $d(\Delta, M)$, and which in addition minimises $d(M, \Pi)$. This helps choosing, among the numerous solutions to the genome halving problem, the one that is closest to the outgroup. We do not study this variant here, and it is open for all genomic distances.

9.2 Results

9.2.1 Breakpoint distance, circular and mixed genomes

In this section, $d = d_{BP}$, and genomes are considered in their most general definition, that is, multichromosomal with both circular and linear chromosomes allowed. All the results also stand for circular genomes, but not always for linear genomes, which will be considered in a following section. As the nuclear genome of a eukaryotic species, a mixed karyotype is rarely observed, so probably unstable. Nevertheless this case is of great theoretical interest, as it is the only combination of distance and karyotype where all five problems mentioned in the previous section prove to be polynomially solvable, including the median problem which is hard for almost every other variant. Furthermore, the solutions in this context may suggest approaches for other variants of the problems, as well as providing a rapid bound for other distances, through the Watterson *et al.* bound [220].

Distance and double distance

The distance computation follows directly from the definition, and is easily achievable in linear time.

The double distance computation is also easy: let Π be a genome and Δ be an all-duplicates genome. Let $a(\Pi, \Delta)$ be the sum, for every adjacency xy in Π, of the number of adjacencies among $x1y1$, $x1y2$, $x2y1$, $x2y2$ in Δ. Let $e(\Pi, \Delta)$ be the sum, for every telomere x in Π, of the number of telomeres among $x1$ and $x2$ in Δ.

Then we obtain

$$d(\Pi, \Delta) = 2n - a(\Pi, \Delta) - \frac{e(\Pi, \Delta)}{2}.$$

Indeed, it is a lower bound on the distance, because $a(\Pi, \Delta)$ and $e(\Pi, \Delta)$ are upper bounds on the number of common adjacencies and common telomeres, respectively, between Δ and any $\Pi \oplus \Pi$. This lower bound is attained by constructing $\Pi \oplus \Pi$ in the following way: let xy be a possibly telomeric adjacency in Π (either x or y may be $\circ$ symbols); if $x1y1$ or $x2y2$ is an adjacency in Δ, choose $x1y1$ and $x2y2$ as adjacencies in $\Pi \oplus \Pi$; If $x1y2$ or $x2y1$ is an adjacency in Δ, choose $x1y2$ and $x2y1$ as adjacencies in $\Pi \oplus \Pi$; the two cases are either mutually exclusive if xy is not telomeric, or identical if xy is telomeric, so the assignment is made without ambiguity. For all adjacencies that have not been assigned, assign them arbitrarily.

Median

The following result contrasts with the NP-completeness proofs of almost all median problems in the literature [29,35,153] (see [21,149] for tractability results on some variants). The problem is NP-complete for unichromosomal genomes, that is, when the median genome M is required to be unichromosomal, whether the genomes are linear or circular [29,153], but the multichromosomal case happens to be easier.

Theorem 2. *There is a polynomial time algorithm for the breakpoint median problem for multichromosomal genomes.*

Proof. Let Π_1, Π_2, Π_3 be three genomes on a gene set $\mathcal{G}$ of size n. For any genome M on $\mathcal{G}$, let $s(M) = d(\Pi_1, M) + d(\Pi_2, M) + d(\Pi_3, M)$ be the *median score* of M.

Draw a graph G on the vertex set containing (1) all extremities of genes in $\mathcal{G}$, and (2) one supplementary vertex t_x for every gene extremity x. For any pair of gene extremities x, y, draw an edge xy weighted by the number of genomes, among Π_1, Π_2, Π_3, for which xy

is an adjacency. Then there is an edge between each pair of gene extremities, weighted by 0, 1, 2 or 3. Now for any vertex x, draw an edge xt_x weighted by half the number of genomes, among Π_1, Π_2, Π_3, having x as a telomere. Each edge xt_x is then weighted by 0, $\frac{1}{2}$, 1, or $\frac{3}{2}$. Finally, put an edge of weight 0 between t_x and t_y for all pairs of gene extremities x, y.

Let M be a perfect matching in G. Clearly, the edges joining gene extremities in M define the adjacencies of a genome, which we also call M. The relation between the weight of the perfect matching M and the median score of the genome M is easy to state:

Claim 1. *The weight of the perfect matching M in G is $3n - s(M)$.*

Indeed, for any genome Π_i, $d(\Pi_i, M) = n - (a_i + \frac{e_i}{2})$, where $a_i = a(\Pi_i, M)$ is the number of common adjacencies between M and Π_i, and $e_i = e(\Pi_i, M)$ is the number of common telomeres between M and Π_i. If M and Π_i have a common adjacency or a common telomere, this accounts for 1 or $\frac{1}{2}$, respectively, in the weight of the perfect matching M. So the weight of the matching M is $w(M) = a_1 + a_2 + a_3 + \frac{e_1+e_2+e_3}{2}$, which yields $d(\Pi_1, M) + d(\Pi_2, M) + d(\Pi_3, M) = 3n - w(M)$.

Conversely, any genome M can be extended to a perfect matching M in G such that $s(M) = 3n - w(M)$: construct the matching M by including the edges xy and $t_x t_y$ for each adjacency xy and an edge xt_x for each telomere x.

Claim 1 implies that a maximum weight perfect matching M is a minimum score median genome. As the maximum weight perfect matching problem is polynomial [120], so is the breakpoint median problem. □

If the three genomes in the instance are circular, then it is possible to constrain the result to also be circular by restricting the graph G to the extremities of the genes. Then,

in the same way, a perfect matching gives a circular solution to the median problem. This is not the case for linear genomes, since there is no way to guaranty that no chromosome in an instance is circular.

Note that a generalisation of this algorithm remains valid if the median of more than three genomes is to be computed. The phylogeny problems, both "big" and "small" versions, which also generalise the median problem for three genomes, remain open. The big problem is the search for a Steiner tree in the space of genomes, minimising the sum of the distances on its branches, while in the small problem, presumably easier, the graph-theoretical structure of the tree, namely its vertex set and edge or branch set, are given, and only the genomes corresponding to the extra vertices (not corresponding to the given genomes) need to be reconstructed.

Halving.

To our knowledge, the genome halving with breakpoint distance has not yet been studied. In this framework, it has an easy solution, using a combination of elements from the maximum weight perfect matching technique in the solution of the median problem presented above, and the double distance computation. Let Δ be an all-duplicates genome on a gene set $\mathcal{G}$, and G be the graph on the vertex set containing (1) all the extremities of the genes in $\mathcal{G}$, and (2) one supplementary vertex t_x for every gene extremity x. For any pair of gene extremities x, y, draw an edge in G weighted by zero, one or two according to the number of adjacencies in Δ among $x1y1$, $x1y2$, $x2y1$,and $x2y2$. Now for any vertex x, draw an edge xt_x weighted by half the number of telomeres among x_1 and x_2 in Δ. Finally, put an edge of weight 0 between $t_x t_y$ for all pairs of gene extremities x, y.

For a genome M on $\mathcal{G}$, define a perfect matching, also called M, by including edges xy and $t_x t_y$ for each adjacency xy, and an edge xt_x for each telomere x. Let $w(M)$ be the weight of the matching M.

Claim 2. *For a genome M on $\mathcal{G}$, the perfect matching M thus constructed satisfies $w(M) = 2n - d(\Delta, M)$.*

Indeed, the score of the perfect matching M is $a(\Delta, M) + \frac{e(\Delta,M)}{2}$, that is, $2n - d(\Delta, M)$, according to the double distance formula (see above in this section).

Conversely, it is easy to see that any perfect matching on G defines a genome M such that $w(M) = 2n - d(\Delta, M)$. This implies that the maximum weight perfect matching solves the genome halving problem in the breakpoint distance context.

Again, it is possible to solve the problem on only circular genomes by restricting the graph G to the gene extremities, dropping the t_x supplementary vertices.

Guided Halving.

As is the case for the median problem, this context provides the only polynomial result for the guided genome halving problem up to our knowlegde. The solution combines elements of the three previous results, on the double distance, median and halving problems.

Let Δ be an all-duplicates genome on a gene set $\mathcal{G}$, and Π be an ordinary genome on $\mathcal{G}$. Let G be the graph on the vertex set containing (1) all the extremities of the genes in $\mathcal{G}$, and (2) one supplementary vertex t_x for every gene extremity x.

For any pair of gene extremities x, y, there is an edge in G weighted by the number of adjacencies among $x1y1$, $x1y2$, $x2y1$, $x2y2$ in Δ, and xy in Π. Now there is an edge xt_x for any gene extremity x weighted by half the number of telomeres among $x1$, $x2$ in Δ

and x in Π. So each edge between gene extremities has an integer weight in $\{0, 1, 2, 3\}$, and xt_x edges may have weight 0, $\frac{1}{2}$, 1, or $\frac{3}{2}$. Add 0-weight edges $t_x t_y$ for all pairs x, y of gene extremities.

For any genome M, let $s(M) = d(\Delta, M) + d(M, \Pi)$. It is possible to construct a perfect matching M in G from genome M by choosing edges xy and $t_x t_y$ for every adjacency xy in M. Its weight is denoted $w(M)$.

Claim 3. *For a genome M, the perfect matching thus constructed satisfies $w(M) = 3n - s(M)$.*

Indeed, the weight of the perfect matching M is $w(M) = a(M, \Pi) + a(M, \Delta) + \frac{e(M,\Pi)+e(M,\Delta)}{2}$. According to the double distance formula (see above in this section), this yields $w(M) = 3n - s(M)$.

Conversely, if M is a perfect matching in G, its edges between gene extremities define the adjacencies of a genome M which satisfies $s(M) = 3n - w(M)$. This implies that the maximum weight perfect matching solves the guided genome halving problem in the breakpoint distance context.

As is the case for the median problem, it is possible to generalise this statement for an arbitrary number of ordinary outgroup genomes. The phylogenetic problems are open.

Again, we can solve the problem on circular genomes by dropping the t_x supplementary vertices in the graph G.

9.2.2 Breakpoint distance, linear case

In this section, $d = d_{BP}$ and all genomes must be linear, as is most appropriate for modeling for the eukaryotic nuclear genome. In contrast to the model of the previous section, all the problems concerning at least three genomes are NP-complete.

Distance and double distance

The solutions to these problems are the same as in the previous section, where circularity was allowed.

In the double distance computation, it is guaranteed that $\Pi \oplus \Pi$ is linear if Π is linear, because if x is a telomere in Π, then both $x1$ and $x2$ are telomeres in $\Pi \oplus \Pi$.

Median

Whereas the median is polynomial in the circular and mixed cases, it changes complexity as soon as median genomes are required to be linear. This does not prevent the use of the polynomial algorithm described above as a lower bound, but all biologically relevant median problems seem in fact to be NP-complete.

Theorem 3. *The breakpoint median problem for multichromosomal linear genomes is NP-hard.*

Proof. We use a reduction from the *2-chromosome breakpoint median*, for which NP-hardness is proved in Lemma 4.

The *2-chromosome breakpoint median* problem takes as input three unichromosomal linear genomes Π_1, Π_2, and Π_3 on a set $\mathcal{G}$ of genes, all having the same pair of telomeres.

It asks for a linear genome M on $\mathcal{G}$ with at most two chromosomes, which minimises its median score $s(M) = d(\Pi_1, M) + d(\Pi_2, M) + d(\Pi_3, M)$.

The following lemma states the difficulty of the breakpoint median problem compared to the 2-chromosome breakpoint median problem, and thus, together with Lemma 4, proves Theorem 3.

Lemma 3. *Let Π_1, Π_2, Π_3, be three unichromosomal linear genomes Π_1, Π_2, and Π_3 all having the same pair of telomeres on the gene set $\mathcal{G} = \{1, \ldots, n\}$, and k be a positive integer. There exists a genome M on $\mathcal{G}$ with at most two linear chromosomes such that $s(M) \leq k$ if and only if there exists a multichromosomal linear genome M' on $\mathcal{G}$ with $s(M') \leq k$.*

($\Rightarrow$): This direction is trivial : simply take $M' = M$. ($\Leftarrow$): Let M' be a linear multichromosomal genome satisfying $s(M') \leq k$, that has as few chromosomes as possible. We will prove that M' has at most two chromosomes. Suppose M' has at least three chromosomes. Then it has at least six telomeres $v_1, \ldots, v_6$. Among them, it is possible to identify two telomeres (say without loss of generality v_1 and v_2), that belong to different chromosomes and are not telomeres in Π_1, Π_2, or Π_3, because by hypothesis, they all have the same two telomeres. Then the genome constructed from M' by adding the adjacency v_1v_2 has at most the same median score as M' and fewer chromosomes, contradicting the hypothesis on M'. So choosing $M = M'$ gives a genome with at most two chromosomes such that $s(M) \leq k$. □

We now need to prove the NP-hardness of the *2-chromosome breakpoint median* problem. We use a reduction from the hamiltonian cycle problem for directed graphs with vertex degree at most three, similar to the proof of Bryant [29] for the breakpoint median problem

for unichromosomal circular genomes.

Lemma 4. *The* 2-chromosome breakpoint median *problem is NP-hard.*

Proof. Given a directed graph with maximum degree 3, deciding if it has a hamiltonian directed cycle is an NP-complete problem [57]. Let thus G_0 be such a digraph. We will construct an instance of the 2-chromosome breakpoint median problem from G_0.

First, let G be the directed graph with vertex set $V(G) = V(G_0) \cup \{x_e \ : \ e \in E(G_0)\} \cup \{p, q\}$, and arc set

$$E(G) = \{ux_{uv}, x_{uv}v \ : \ uv \in E(G_0)\} \cup \{pq\}.$$

Note that G_0 has a hamiltonian cycle if and only if there is a cycle in G covering all vertices but p and q.

Given a subset $X \subseteq E(G)$ of the arcs of a graph G, let G_X denote the graph with vertex set $V(G)$ and arc set X.

Construct three subsets A, B, C of arcs of G such that every arc of G belongs to exactly one of A, B, C, and A, B, C are either hamiltonian cycles of G or sets of vertex disjoint paths in G. The procedure is straightforward: for all vertex v of G which is a vertex of G_0, put all incoming arcs in different subsets, and all outgoing arcs in different subsets. It can be done independently for every vertex because from the construction of G, no two vertices of G_0 are neighbors. Eventually put arc pq in any subset.

Now perform a series of modifications of G to obtain a graph with three hamiltonian cycles A, B, and C. Along these modifications, we maintain two subsets of arcs called respectively *supplementary arcs* and *original arcs*, noted S and O. Before any transformation,

S is empty and O contains all arcs of G. The goal is to maintain the property that there exists a cycle covering all vertices of G except p and q and using only original arcs if and only if there is a hamiltonian cycle in G_0. As already remarked, the property is true at the beginning. Choose $X \in \{A, B, C\}$ such that X is not a hamiltonian cycle in G (it is a set of disjoint paths). Choose two vertices a and b such that adding the arc ab to G and X would either give a graph in which X is a hamiltonian cycle, or a set of disjoint paths with fewer components. Then choose any vertex x of G different from a and b. Perform the following transformation illustrated in Figure 4: add two new vertices y and z. For each arc xw of G, replace it by the arc zw, and add zw to $Y \in \{A, B, C\}$ whenever $xw \in Y$, and to O whenever $xw \in O$. Add arcs xy, yz, xz, ay, yb to G. Add xz, ay, yb to X and to S. Also add xy, yz to all $\{A, B, C\} \setminus X$.

Repeat this process until A, B and C are all hamiltonian cycles in the resulting graph, which we call G'. The *weight* of an arc of G' is the number of hamiltonian cycles among A, B and C which contain this arc. Note that G' has only weight 1 and 2 arcs.

Let v be an arbitrary vertex of G, different from p and q. Let $\mathcal{G} = V(G') \setminus \{v\} \cup \{v_1, v_2\}$ be a set of genes (v_1 and v_2 are two new genes). For every $X \in \{A, B, C\}$, construct a genome Π_X on $\mathcal{G}$ such that $x^h y^t$ is an adjacency in genome X if xy is an arc of X in G' and x, y are different from v; If xv and vy are the arcs of X covering v, add the adjacencies $x^h v_1^t$ and $y^t v_2^h$. This gives three linear unichromosomal genomes on $\mathcal{G}$ with the same pair of telomeres (v_1^h and v_2^t), thus an instance of the *2-chromosome breakpoint median* problem. For any genome M on $\mathcal{G}$, the *weight* of an adjacency xy is the number of genomes, among Π_A, Π_B and Π_C, which contain this adjacency. Adjacencies $u^h w^t$ in M such that uw is in S arc are called *supplementary adjacencies.*

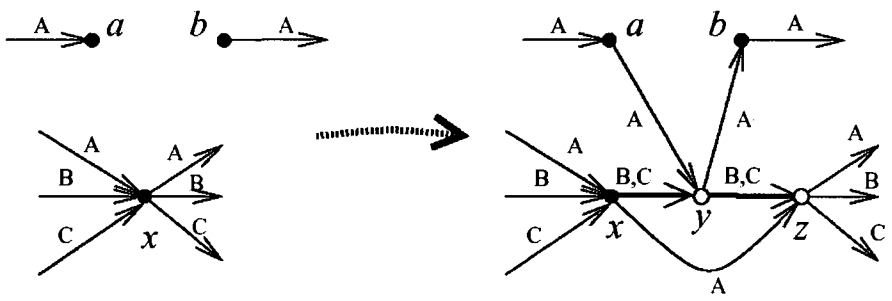

Figure 9.4: Reduction of hamiltonian cycle to linear breakpoint median. This figure is redrawn from [29].Vertex a has no outgoing arc with $X = A$ in its label set, and b has no incoming arc with A in its label set. We choose a, b such that adding arc ab to $G[A]$ would not give a non-Hamiltonian circuit. We choose an another vertex x and insert two new vertices y and z. The incoming arcs of x in the right hand graph are the same as in the left hand graph. The outgoing arcs of z are the same as the incoming edges of x in the left hand graph. The remaining edges reduce the number of components in $G[A]$ but leave the same number of components in $G[B]$ and $G[C]$.

Let λ_i be the number of arcs of weight i in the graph G', for each $0 \leq i \leq 3$. For any genome M on $\mathcal{G}$, note $s(M) = d(\Pi_A, M) + d(\Pi_B, M) + d(\Pi_C, M)$. The following is inspired by a result from [169] used in [29].

Claim 4. *Let $n = |V(G')|$. A genome M with N_M linear chromosomes on $\mathcal{G}$ satisfies $s(M) \geq 2n - 1 + N_M - \lambda_2$, where equality holds if and only if M contains all adjacencies of weight 2, and no adjacency of weight 0.*

Indeed, for a genome M, denote by $w(xy)$ the weight of the adjacency xy, and $l_i = |xy$ adjacency of $M \ : \ w(xy) = i|$, for each $0 \leq i \leq 3$. Let $t_1 = 1$ if M has v_1^h as a telomere, and $t_1 = 0$ otherwise, and $t_2 = 1$ if M has v_2^t as a telomere, and $t_2 = 0$ otherwise. Then we may write $s(M) = 3|\mathcal{G}| - (3l_3 + 2l_2 + l_1 + 3\frac{t_1+t_2}{2})$. As genome M has N_M chromosomes, we have $l_3 + l_2 + l_1 + l_0 = |\mathcal{G}| - N_M$ and $|\mathcal{G}| = n + 1$, so $s(M) = 2n + 2 + N_M - 2l_3 - l_2 - 3\frac{t_1+t_2}{2} + l_0 \geq 2n - 1 + N_M - 2\lambda_3 - \lambda_2$. As by construction no arc of G' has weight 3, we may write $s(M) \geq 2n - 1 + N_M - \lambda_2$. Equality holds if and only if $l_2 = \lambda_2$, $l_0 = 0$, and $3\frac{t_1+t_2}{2} = 3$, that is, if and only if M contains all adjacencies of weight at least 2, and no adjacency of weight 0, because $l_0 = 0$ implies that v_1^h and v_2^t are telomeres of M, thus $3\frac{t_1+t_2}{2} = 3$.

Claim 5. *There is a linear genome M on $\mathcal{G}$ with at most two chromosomes, with $s(M) = 2n - 1 + N_M - \lambda_2$ if and only if there is a hamiltonian cycle in G_0.*

($\Rightarrow$) Suppose there is a linear genome M on $\mathcal{G}$ with at most two chromosomes, with $s(M) = 2n - 1 + N_M - \lambda_2$. This implies by Claim 4 that M contains all adjacencies of weight 2, no adjacency of weight 0, and that v_1^h and v_2^t are telomeres of M. From the construction of G', M cannot contain any supplementary adjacency, since the extremities

of supplementary adjacencies all are also extremities of weight two adjacencies, which are all contained in M. Note that in G', paths between vertices p and q to other vertices of the graph necessarily contain supplementary arcs. This yields that M has two chromosomes, one containing gene extremities from p and q, and the other containing the gene extremities from the other vertices, with telomeres v_1^h and v_2^t. Let $H \subseteq E(G)$ contain the original arcs $xy \in O$ such that $x^h y^t$ is an adjacency in M, plus the arcs xv and vy for adjacencies of type $x^h v_1^t$ and $v_2^h y$ in M. $H \setminus \{pq\}$ yields a hamiltonian cycle in G_0.

($\Leftarrow$) Suppose there is a hamiltonian cycle in G_0. Then there is a cycle H covering all vertices of G' except p and q. Construct genome M on $\mathcal{G}$ by adding (1) adjacencies $x^h y^t$ whenever xy is in H and x, y are different from v (2) adjacencies $x^h v_1^t$ and $v_2^h y$ whenever xv or vy are arcs of H, (3) all weight two adjacencies, and (4) the adjacency $p^t q^h$. It is easy to check that M is a genome on $\mathcal{G}$, and by construction it contains all weight 2 adjacencies and no weight 0 adjacencies.

This proves that it is NP-complete to decide if a median genome reaches the lower bound of Claim 4 for its score, thus computing the minimum score median genome is NP-hard. □

A byproduct of this proof is the NP-hardness of the 2-chromosome breakpoint median problem. The result implies NP-completeness of the general case, where the genomes in the instance do not necessarily have the same pairs of telomeres. A consequence is that for any fixed k, it is also NP-hard to compute a best linear median genome with at most k chromosomes.

Halving.

Surprisingly, this problem has not been treated in the literature. We conjecture it has a polynomial solution, because the halving problem for all other rearrangement distances is polynomial. Constructing a solution is beyond the scope of this paper, and the problem remains open.

Guided Halving.

This problem is NP-hard, as proved in [241], using the NP-completeness result for the median proved just above in this section.

9.2.3 DCJ distance, general case

In this section, $d = d_{DCJ}$. Genomes can have several chromosomes, circular or linear. This is the most general context in which the DCJ distance has been explicitly formulated [17]. Genomes rarely contain both circular and linear chromosomes; eukaryote nuclear genomes contain multiple linear chromosomes while prokaryotes generally contain one large circular chromosome, sometimes with additional *plasmids*. Nevertheless, the simplicity of the computational framework where both genomes may contain both circular and linear chromosomes makes it attractive to mathematical study.

Note that the complexity of the median problem is not established by the work of Caprara [35], who proved the unichromosomal result only. We show the NP-hardness of the multichromosomal case here.

Alekseyev and Pevzner [7] mention that the complexity of the double distance problem in the context of circular genomes is open; we show here that it is NP-hard as well.

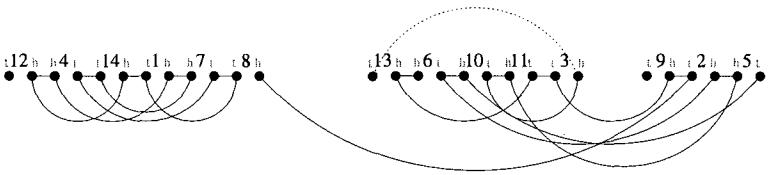

Figure 9.5: A Breakpoint Graph. The breakpoint graph of the genomes Π (see Figure 9.1) and Γ, given by the union of
$C_1 = \{T12^t, 12^h14^h, 14^t7^h, 7^t4^t, 4^h1^h, 1^t8^t, 8^h2^t, 2^h6^t, 6^hT\}$ and
$C_2 = \{T9^t, 9^h3^t, 3^h10^t, 10^h5^t, 5^h11^h, 11^t13^h, 13^tT\}$. Π-edges are dotted lines, and Γ-edges are plain lines.

Distance.

There is an easy linear solution, both for the distance and the scenario computation [17,231]. We briefly recall the formula for computing the distance, because the underlying principle will be used in our proofs later on in this paper.

The *breakpoint graph* of two genomes Π and Γ on a gene set $\mathcal{G}$, denoted by $BP(\Pi, \Gamma)$, is the graph whose vertex set is the set of extremities of the genes in $\mathcal{G}$, where there is an edge between two vertices x and y if xy is an adjacency in either Π (these are Π-edges) or Γ (Γ-edges). Note that we do not invoke any $\circ$ symbols in the construction of the breakpoint graph. Vertices in this graph have degree zero, one or two, so that the graph is a set of paths (possibly including some with no edges) and cycles. It is also the line-graph of the *adjacency graph*, an alternate representation in [17]. Figure 9.5 shows an example of a breakpoint graph. Theorem 4 shows how to obtain the distance directly from the graph.

Theorem 4. *[17][4] For two genomes Π and Γ on a gene set $\mathcal{G}$ of size n, let $c(\Pi, \Gamma)$ be the number of cycles of the breakpoint graph $BP(\Pi, \Gamma)$, and $p(\Pi, \Gamma)$ be the number of paths with*

[4]The formula is presented in [17] with the cycles and odd paths of the adjacency graph. This corresponds to cycles and even paths of the breakpoint graph, as it is the line-graph of the adjacency graph.

an even number of edges. Then

$$d(\Pi, \Gamma) = n - c(\Pi, \Gamma) - \frac{p(\Pi, \Gamma)}{2}.$$

Note the similarity to the breakpoint distance formula in the background section on page 219. The number of genes n is the same in both formulae, the parameter c is related to parameter a in the breakpoint formula in that each common adjacency is a cycle of the breakpoint graph (with two parallel edges), and parameter p is related to parameter e, as each shared telomere is an even path (with no edge) in the breakpoint graph. Although these two measures of genomic distance were derived in different contexts and through different reasoning, their formulae show a remarkably similar form. They differ in that the DCJ formula also counts non-trivial cycles and paths, but for distant genomes, both measures tend to give similar values.

Double distance.

The NP-completeness proof for the double distance problem follows the principles of Caprara's hardness proof for the median problem in the unichromosomal case [35].

Theorem 5. *The DCJ double distance problem is NP-hard for multichromosomal mixed or circular genomes.*

Proof. The reduction is from the *breakpoint graph decomposition* (BGD) problem (see [35]). A graph G is *bicoloured* if all its edges are coloured either red or blue; it is *balanced* if it has only degree 2 or degree 4 vertices, every vertex is incident to the same number of red and blue edges, and there is no cycle formed by only red or only blue edges. Given a balanced

bicoloured graph G, the breakpoint graph decomposition problem is to find a partition of the edges of G into a maximum number of edge-disjoint cycles, each alternating between red and blue edges. Caprara [35] first proved the NP-hardness of this problem, and Berman and Karpinski [20] extended this by proving APX-hardness.

Let G be a balanced bicoloured graph on n vertices, defining an instance of the BGD problem. Let $w2$ be the number of degree 2 vertices of G, and $w4$ be the number of degree 4 vertices of G. Define the gene set $\mathcal{G}$ as the vertex set of G. Construct an all-duplicates genome Δ and a genome Π on $\mathcal{G}$ in the following way, as illustrated in Figure 6. First, for each gene X of $\mathcal{G}$, let X^tX^h be an adjacency in Π. Then, for every vertex X of G, let $X1^t$, $X1^h$, $X2^t$ and $X2^h$, be the extremities of the duplicated gene X. If X has degree two in G, add the adjacency $X1^tX2^h$ in Δ (if X has degree four, no adjacency is added at this point). Then for each blue edge XY in G, choose among $X1^h$ and $X2^h$ an extremitiy that is not yet involved in an adjacency, and another among $Y1^h$ and $Y2^h$ (arbitrarily if neither is involved in an adjacency yet). Add an adjacency between the two chosen extremities in Δ. Then for each red edge XY in G, choose among $X1^t$ and $X2^t$ an extremitiy that is not yet involved in an adjacency, and another among $Y1^t$ and $Y2^t$ (arbitrarily if neither is involved in an adjacency yet). Add an adjacency between the two chosen extremities in Δ.

We then have an all-duplicates genome Δ, and a genome Π. Note that Π is composed of n circular chromosomes, one for each gene, and that neither Π nor Δ have telomeres.

Claim 6. *The maximum number of edge-disjoint alternating cycles in G is equal to $2n - d(\Delta, \Pi) - w2$.*

(This claim implies the theorem).

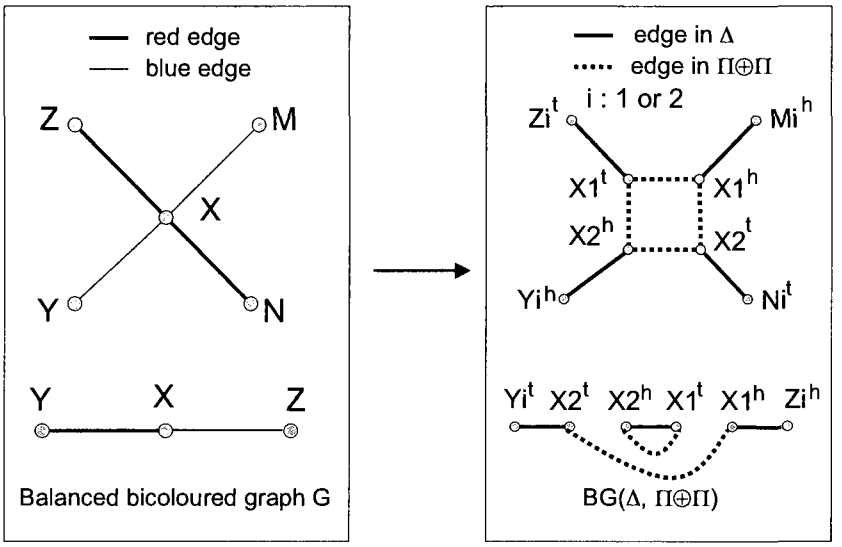

Figure 9.6: Reduction of BGD to DCJ double distance problem,The left hand graph is the balanced bicoloured graph G, and the right hand graph represents the adjacencies of the constructed genomes Δ and $\Pi \oplus \Pi$.

We first prove that the maximum number of edge-disjoint alternating cycles in G is at least $2n - d(\Delta, \Pi) - w2$. Let $\Pi \oplus \Pi$ be the doubled genome such that $d(\Delta, \Pi \oplus \Pi) = d(\Delta, \Pi)$. As no genome has a telomere, by Theorem 4, $d(\Delta, \Pi \oplus \Pi) = 2n - c(\Delta, \Pi \oplus \Pi)$. Therefore there are $c(\Delta, \Pi \oplus \Pi)$ edge-disjoint cycles in $BG(\Delta, \Pi \oplus \Pi)$ alternating between $\Pi \oplus \Pi$-edges and Δ-edges. Among them, $w2$ cycles are containing only two edges : if a vertex X of G has degree 2, then $\Pi \oplus \Pi$ has the adjacency $X1^t X2^h$ and $X2^t X1^h$ because the other possibility systematically has one cycle less in $BG(\Delta, \Pi \oplus \Pi)$. The Δ edges of all the other cycles are the edges of an alternating blue-red cycle in G. Indeed, every blue edge defines an adjacency in Δ containing two gene heads, and every red edge defines an adjacency containing two gene tails. The $\Pi \oplus \Pi$-edges all join one tail and one head, so two consecutive Δ-edges in a cycle of $BG(\Delta, \Pi \oplus \Pi)$ have different colours. This means there are at least $2n - d(\Delta, \Pi) - w2$ alternating cycles in G.

Conversely, if there are k edge-disjoint alternating cycles in G, then $d(\Delta, \Pi) \leq 2n - k - w2$. Indeed, let C be any cycle of this partition. For every covered vertex X of degree 4 in G, let e and f be two consecutive edges of C (say e is blue and f is red) which are both incident to X. If e defines an adjacency in Δ which contains $X2^h$ and f defines an adjacency which contains $X2^t$, choose $X1^hX1^t$ and $X2^hX2^t$ as adjacencies for $\Pi \oplus \Pi$. If e defines an adjacency which contains $X2^h$ (or $X1^h$) and f defines an adjacency which contains $X1^t$ (or $X2^t$), choose $X1^hX2^t$ and $X2^hX1^t$ as adjacencies for $\Pi \oplus \Pi$. For vertices of degree 2, always choose $X1^hX2^t$ and $X2^hX1^t$ as adjacencies for $\Pi \oplus \Pi$. In this construction, each red-blue alternating cycle in G is a $\Pi \oplus \Pi$-Δ alternating cycle in $BP(\Pi \oplus \Pi, \Delta)$ that has at least k cycles. And there are $w2$ additional length 2 cycles at each degree 2 vertex. So $d(\Delta, \Pi) \leq 2n - k - w2$. □

Median

Though effective exact algorithms [230] and heuristics [1, 107] are available, we have:

Theorem 6. *The DCJ median problem for multichromosomal genomes is NP-hard, even for circular genomes.*

Proof. We use a reduction from the breakpoint graph decomposition defined in the proof of Theorem 5, in a way very similar to part of Caprara's proof [35] for the unichromosomal case.

Let G be a balanced bicoloured graph on n vertices. Define the gene set $\mathcal{G}$ as a set containing one gene X for every degree 2 vertex of G, and two genes X and Y for every degree 4 vertex of G.

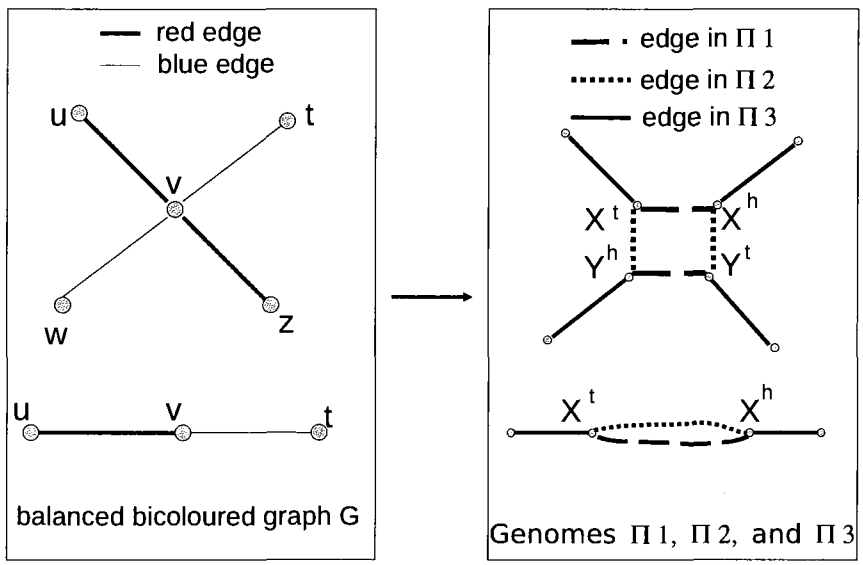

Figure 9.7: Reduction of BGD to DCJ median problem,The left hand graph is the balanced bicoloured graph G, and the right hand graph represents the adjacencies of the constructed genomes Π_1, Π_2 and Π_3.

Then construct the genomes Π_1, Π_2, Π_3 in the following way, which is similar to the transformation in [35], as illustrated in Figure 9.7.

For each degree 4 vertex v of G, add the two adjacencies X^tX^h and Y^tY^h to Π_1, and the two adjacencies X^tY^h and Y^tX^h to Π_2.

Now fo every blue edge vw in G, add to Π_3 an adjacency between heads of genes contructed from v and w (choose one duplicated arbitrarily if v or w have degree 4). And for every red edge vu in G, add to Π_3 an adjacency between tails of genes contructed from v and u (choose one duplicated arbitrarily if v or u have degree 4).

It is easy to see that Π_1, Π_2, and Π_3 define genomes on the set of genes $\mathcal{G}$, and they have no telomeres. Let $w2$ be the number of degree 2 vertices of G, and $w4$ be the number

of degree 4 vertices of G.

Claim 7. *There exists a genome M on $\mathcal{G}$ such that $d(M,\Pi_1) + d(M,\Pi_2) + d(M,\Pi_3) \leq w2 + 3w4 - k$ if and only if there exists at least k edge-disjoint alternating cycles in G.*

(This claim implies the theorem.)

($\Leftarrow$): Suppose there are k edge-disjoint alternating cycles in G. We will construct a median genome M such that $d(\Pi_1, M) + d(\Pi_2, M) + d(\Pi_3, M) = w2 + 2w4 - k$. First, for each degree 2 vertex v of G, let X^tX^h be in M. Then, let v be a degree 4 vertex in G, and vw be a blue edge incident to v. In an alternating cycle, edge vw is consecutive with a red edge, say uv. To vw is associated a constructed Π_3 adjacency, say X^hW^h. Then either the Π_3 adjacency associated to uv contains the gene extremity X^t, or it contains the extremity Y^t. In the first case, let X^hX^t and Y^hY^t be in M, and in the second case, let X^hY^t and X^tY^h be in M.

The matching M defines the adjacencies of a circular genome on $\mathcal{G}$, that we also call M. There are $w2+2w4$ genes, so by Theorem 4, $d(\Pi_1, M)+d(\Pi_2, M)+d(\Pi_3, M) = 3(w2+2w4) - (c(\Pi_1, M)+c(\Pi_2, M)+c(\Pi_3, M))$. By construction, we have $c(M,\Pi_1)+c(M,\Pi_2) = 2w2 + 3w4$, and $c(M,\Pi_3) = k$, so $d(\Pi_1, M) + d(\Pi_2, M) + d(\Pi_3, M) = w2 + 3w4 - k$.

($\Rightarrow$): Suppose M is a genome such that $d(\Pi_1, M)+d(\Pi_2, M)+d(\Pi_3, M) \leq w2+3w4-k$. Suppose M is chosen such that $d(\Pi_1, M) + d(\Pi_2, M) + d(\Pi_3, M)$ is minimum, and among all such genomes, choose M with a maximum number of edges parallel to Π_1-edges or Π_2-edges. A circular genome is said to be *canonical* if it only has adjacencies that belong to Π_1 or Π_2. We will prove that M is canonical.

Subclaim. M is canonical.

Suppose M is not canonical. Suppose first that there is a degree 2 vertex v in G, such that M does not contain the adjacency X^tX^h. Suppose M contains adjacencies X^ta and X^hb, where a and b are gene extremities or $\circ$ symbols if X^t or X^h is a telomere in M. Then replace X^ta and X^hb by X^hX^t and ab (simply X^hX^t if both X^t and X^h are telomeres in M). By this operation, $c(M,\Pi_1)$ and $c(M,\Pi_2)$ both increase by at least 1, and $c(M,\Pi_3)$ decreases by at most 1, so $d(\Pi_1,M)+d(\Pi_2,M)+d(\Pi_3,M)$ decreases by one, contradicting the hypothesis.

Now suppose that there is a degree 4 vertex in G, such that M does not contain any of the adjacencies X^hX^t, Y^hY^t, X^hY^t, Y^hX^t. Say it contains adjacencies X^ha, X^tb, Y^hc, Y^td, where a,b,c,d may be null symbols if any of X^h, X^t, Y^h, Y^t is a telomere in M. Then replace X^ha, X^tb, Y^hc, Y^td by X^hX^t, Y^hY^t, and either ab, cd, or ac, bd, or ad, bc, according to the combination that creates the largest number of cycles in $BP(M,\Pi_3)$. Suppose now that M contains only one among the adjacencies X^hX^t, Y^hY^t, X^hY^t, Y^hX^t, say X^hX^t, and M has adjacencies Y^tb and Y^hc. Then replace edges Y^tb and Y^hc by Y^hY^t and bc. All these operations decrease $d(\Pi_1,M)+d(\Pi_2,M)+d(\Pi_3,M)$ or maintain it constant, while increasing the number of edges parallel to Π_1 and Π_2, contradicting the hypothesis. So the subclaim is proved.

Now, since M is canonical, there are $c(\Pi_3,M)$ edge-disjoint alternating cycles in G, since an adjacency of M always joins a head and a tail, so the corresponding edge in G is adjacent to one red edge at one of its vertices and one blue edge at the other. By Theorem 4, $c(\Pi_3,M) = 3(w2+2w4)-(d(\Pi_1,M)+d(\Pi_2,M)+d(\Pi_3,M)+c(\Pi_1,M)+c(\Pi_2,M))$ and, by hypothesis, $c(\Pi_3,M) \geq 3(w2+2w4)-(w2+2w4-k+2w2+3w4)$, that is, $c(\Pi_3,M) \geq k$, which proves the claim. □

Halving.

This problem has a polynomial solution, as recently stated for unichromosomal genomes by [7] and in the general case by [136, 219]. All these algorithms are simplified versions of the algorithm by El-Mabrouk and Sankoff [51], developed for the RT rearrangement distance, which allows reversals, translocations, fusions and fissions, but not the other DCJ operations.

Guided Halving.

Theorem 7. *The DCJ guided halving problem is NP-complete for multichromosomal genomes.*

Proof. Again, we use a reduction of the breakpoint graph decomposition problem, as in the proofs of Theorems 5 and 6.

Let G be a balanced bicoloured graph on n vertices. Define the gene set $\mathcal{G}$ as a set containing one gene X for every degree 2 vertex of G, and two genes X and Y for every degree 4 vertex of G. From G, we define one genome Π and one all-duplicates genome Δ on $\mathcal{G}$ as illustrated in Figure fig:reduGGH.

For every degree 2 vertex v of G, let X be the corresponding gene in $\mathcal{G}$ and X^t and X^h its extremities. For every degree 4 vertex v of G, let X and Y be the two genes in $\mathcal{G}$, and X^t and X^h, Y^t and Y^h their extremities. For every blue edge uv in G, construct an adjacency in Π between the heads of the genes constructed from u and v, choosing arbitrarily between the heads of vertices X and Y if u or v have degree 4, in such a way that no two adjacencies share an extremity (choose a different head for the two blue edges incident to a degree 4 vertex). For every red edge tv in G, construct an adjacency in Π between the tails of the

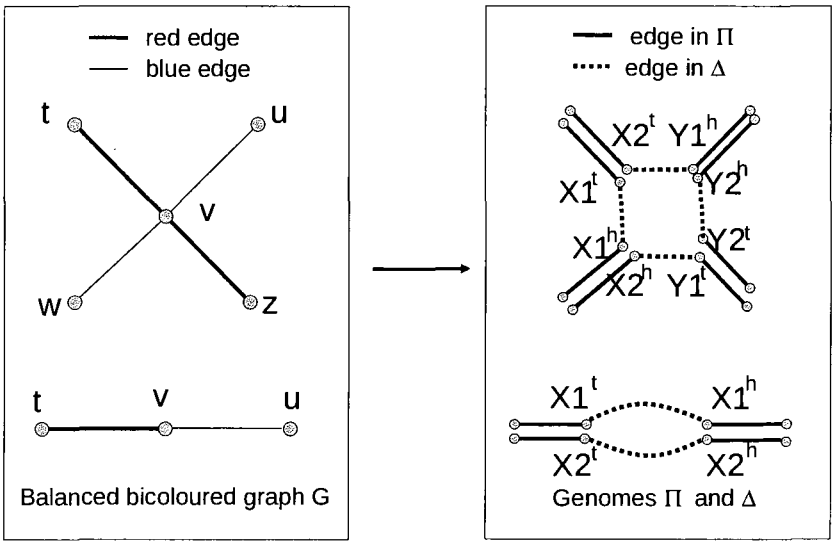

Figure 9.8: Reduction of BGD to DCJ guided halving problem,The left hand graph is the balanced bicoloured graph G, and the right hand graph represents the adjacencies of the constructed genomes Δ and $\Pi \oplus \Pi$.

genes constructed from t and v, again choosing arbitrarily between the tails of vertices X and Y if t or v have degree 4. This defines a genome Π on $\mathcal{G}$ that has no telomere.

Now, define the genome Δ in the following way. For each degree 2 vertex of G, Δ contains the adjacencies $X1^hX1^t$ and $X2^hX2^t$ defined by the extremities of the two copies of gene X. For each degree 4 vertex in G, Δ contains the adjacencies $X1^hX1^t$, $Y2^hY2^t$, $X2^hY1^t$, and $X2^tY1^h$ defined by the extremities of the two copies of the two genes X and Y. This defines an all-duplicates genome Δ on $\mathcal{G}$ that has no telomere. Let $w2$ be the number of degree 2 vertices of G, and $w4$ be the number of degree 4 vertices of G.

Claim. There exists a genome M such that $d(M,\Pi) + d(M,\Delta) \leq w2 + 3w4 - k$ if and only if there exists at least k edges-disjoint alternating cycles in G. (This claim implies the theorem.)

($\Leftarrow$): Suppose there are k edges-disjoint alternating cycles in G. We will construct a genome M and a doubled genome $M \oplus M$ such that $d(M,\Pi) + d(M,\Delta) \leq w2 + 3w4 - k$. First, for each degree 2 vertex v of G, let X^tX^h be in M, and $M \oplus M$ be constructed so that $X1^tX1^h$ and $X2^tX2^h$ are in $M \oplus M$. Then, for every vertex v of degree 4 of G, either the blue edge incident to X^h is followed by the red edge incident to X^t in one alternating cycle, or it is followed by the red edge incident to Y^t. In the first case, let X^hX^t and Y^hY^t be in M, and $X1^hX1^t$, $X2^hX2^t$, $Y1^hY1^t$, $Y2^hY2^t$ be in $M \oplus M$. In the second case, let X^hY^t and X^tY^h be in M, and $X2^hY1^t$, $X1^hY2^t$, $X1^tY2^h$, $X2^tY1^h$ be in $M \oplus M$.

The matching M defines a genome without telomeres, and $M \oplus M$ is one of its doubled genomes.

There are $w2 + 2w4$ genes, so by Theorem 4, $d(\Pi, M) = w2 + 2w4 - c(\Pi, M) = w2 + 2w4 - k$; and $d(\Delta, M \oplus M) = 2(w2 + 2w4) - (2 \times w2 + 3 \times w4)$. So $d(\Pi, M) + d(\Delta, M) =$

$w2 + 3w4 - k$.

($\Rightarrow$): Suppose M is a genome such that $d(\Pi, M) + d(\Delta, M) \leq w2 + 3w4 - k$. Suppose M is chosen such that $d(\Pi, M) + d(\Delta, M)$ is minimum, and among all such genomes, choose M with a maximum number of edges of type X^tX^h for a degree 2 vertex in G, or X^hX^t, Y^hY^t, X^hY^t and X^tY^h for degree 4 vertices of G.

A genome is said to be *canonical* if it has no telomeres and has only edges of type X^tX^h for a degree 2 vertex in G, or X^hX^t, Y^hY^t, X^hY^t and X^tY^h for degree 4 vertices of G. We will prove that M is canonical.

Subclaim. M is canonical.

Suppose M is not canonical. Suppose first that there is a degree 2 vertex v in G, such that M does not contain the edge X^tX^h. Suppose M contains edges X^tu and X^hv, where u and v are gene extremities or $\circ$ symbols if X^t or X^h is a telomere in M. Then replace X^tu and X^hv by X^hX^t and uv (simply X^hX^t is both X^t and X^h are telomeres in M). By this operation, $c(M, \Pi)$ decreases by at most one, while defining $M \oplus M$ as containing $X1^hX1^t$and $X2^hX2^t$ makes $c(M \oplus M, \Delta)$ increase by at least 2. This contradicts the hypothesis.

Now suppose that there is a degree 4 vertex in G, such that M does not contain any of the edges X^hX^t, Y^hY^t, X^hY^t, Y^hX^t. Say it contains edges X^ht, X^tu, Y^hv, Y^tw, where t, u, v, w may be null symbols if any of X^h, X^t, Y^h, Y^t is a telomere in M. Then replace X^ht, X^tu, Y^hv, Y^tw by X^h,X^t, Y^hY^t, tu, vw or X^hY^t, tw, X^tY^h, uv, depending on the cycles in $c(M \oplus M, \Delta)$.

Suppose now that M contains only one among the edges X^hX^t, Y^hY^t, X^hY^t, Y^hX^t say X^hX^t, and M has edges Y^tu and Y^hv. Then replace edges Y^tu and Y^hv by Y^hY^t

and uv. All these operations decrease $d(\Pi, M) + d(\Delta, M)$ or maintain it constant, while increasing the number of edges of type X^tX^h for a degree 2 vertex in G, or X^hX^t, Y^hY^t, X^hY^t and X^tY^h for degree 4 vertices of G, contradicting the hypothesis.

At the end of this process, M is canonical, so the subclaim is proved.

Now, since M is canonical, there are $c(\Pi, M)$ edge-disjoint alternating cycles in G, since an edge of M always joins a head and a tail, so that it is adjacent to one red and one blue edge. By Theorem 4, $c(\Pi, M) = 3(w2 + 2w4) - (d(\Pi, M) + d(\Delta, M \oplus M) + c(\Delta, M \oplus M))$, and by hypothesis, $c(\Pi_3, M) \geq 3(w2 + 2w4) - (w2 + 3w4 - k + 2w2 + 3w4))$, that is, $c(\Pi_3, M) \geq k$, which proves the claim. □

9.2.4 DCJ distance, linear chromosomes

In the original formulation of the DCJ distance [?], it was shown that there is a solution where each excision of a circular intermediate could be followed directly by its reinsertion. Thus the median and halving problems can be stated in terms of exclusively linear chromosomes in both the data genomes and the reconstructed ancestor. They all remain open.

9.2.5 Reversal/Translocation distance

Hannenhalli and Pevzner proposed a polynomial-time algorithm for calculating $d_{RT}(\Pi, \Gamma)$ for two genomes Π and Γ [70]. This was reformulated in [210], minor corrections were added in [148] and [85], and Bergeron *et al.* simplified the formula [19].

A polynomial time genome halving algorithm was given in [51]. Though the constrained DCJ distance in the previous section is arguably just as realistic, because of the long history

of d_{RT}, effective heuristics for RT have been developed and applied for the double distance [242,244], median [26,107] and guided halving problems [238,242,244], but their complexities remain open questions. Note that [38] gives an NP-completeness result on a problem which slightly generalizes the reversal double-distance probem on unichromosomal genomes.

9.3 Conclusions

Table 9.1 summarises the current knowledge of the complexity of the five genome rearrangement problems, including the new results in this paper. Note that all the results on general multichromosomal genome (that is, circular or linear) also hold for exclusively circular genomes, as the polynomial algorithms can always provide a circular solution to a circular instance, and all NP-completeness proofs are constructed with circular chromosomes.

Acknowledgements

Research supported in part by a grant to DS and a doctoral fellowship to CZ from the Natural Sciences and Engineering Research Council of Canada (NSERC). ET is funded by the Agence Nationale pour la Recherche (ANR-08-GENM-036-01 and NT05-3_45205) and the Centre National de la Recherche Scientifique (CNRS). DS holds the Canada Research Chair in Mathematical Genomics. Figures 19.1, 29.2, 39.3 and 59.5 are the work of Sèverine Bérard, used with her kind permission.

Table 9.1: Status of complexity questions for five problems related to ancestral genome reconstruction, for eight genomic distances in the unichromosomal and multichromosomal contexts.Note that unichromosomal problems require that both input and output genomes be unichromosomal, so all problems involving doubled genomes are computationally defined in the circular case, when the doubled genome consists in a single circular chromosome composed of two successive occurences of the ordinary genome. Other versions of the halving problem are less restrictive [7, 51, 219]. P and NP stand for polynomial and NP-hard, respectively, and when followed by ?, represent our conjectures.

problem context: distance, #chr, linear, circular or mixed	distance	halving	double distance	median	guided halving
breakpoint unichr, circular or linear	P	open	open	NP [29,153]	open
breakpoint multichr, circular and mixed	P new	P new	P new	P new	P new
breakpoint multichr, linear	P new	open P?	P new	NP new	NP [241]
DCJ unichr, circular or linear	P [17,231]	P [7]	open	NP [35]	open
DCJ multichr, circular and mixed	P [17,231]	P [136,219]	NP new	NP new	NP new
DCJ multichr, linear	P [231]	open	open	open NP?	open NP?
RT unichr	P [71]	open	open	NP [35]	open
RT multichr	P [70,85,148,210]	P [51]	open NP?	open NP?	open NP?

Chapter 10

Conclusions and future work

10.1 Conclusions

In this thesis I presented several algorithms for use in constructing genome rearrangement phylogenies, where some of the input genomes may be descendants of whole genome doubling events. Throughout, I tested my methods on real data bearing on cereal, yeast and angiosperm genomes. Many ancillary problems emerged along the way and I explored these as well, using both theoretical and applied approaches.

10.1.1 Uniqueness

The original genome halving problem can generate very many equally valid solutions, and these solutions may be very different from each other. Eliminating, or at least reducing, this non-uniqueness was the motivation for starting this work. Using outgroup or reference genomes to guide genome halving can be effective in reducing non-uniqueness drastically. My first method was to explore all the possible genome halving solutions to find the one that

minimizes the sum of two distances: the distance between the inferred tetraploid ancestor and the outgroup, and the distance between this same ancestor and its modern descendant. For the maize data which has 68 syntenic blocks, I obtained 8 best guided genome halving solutions out of $1,500,000$ different unconstrained solutions by using sorghum as the outgroup and 24 best genome halving solutions by using rice as the outgroup. After doing this, I could almost completely eliminate the non-uniqueness by a local search for the optimized ancestor not constrained to be the genome halving solution, deriving only two solutions for maize-sorghum data and one unique solution for maize-rice data.

While this "brute force" method works well with maize, it is not computationally feasible for yeast and *Populus* data which consist of many more duplicated genes. (≥ 518 for yeast, ≥ 1955 for *Populus*). In my first effort to get around this problem, I devised a strategy for obtaining diverse samples of the search path, hoping to cover enough of the solution space such that a local optimization applied to all the best samples would hit upon the global optimum. This seemed to work fairly well for the yeast data although the local optimization turned out to be a serious rate-limiting factor. I was thus motivated to abandon the three-step procedure and to integrate consideration of the outgroup into the halving algorithm itself. A single run of the algorithm thus replaces a census of all possible halving solutions or of a large sample of solutions. This achieves a tremendous saving of computational time, even though the linear running time of the halving algorithm is compromised (becoming worst-case cubic). Table 5.1 shows the performances of the first and the improved versions of GGH algorithms. The new algorithm, characterized by the data structure I call "path groups", is still a heuristic. Even though it takes account of the outgroup genome in constructing a halving solution, it does not necessarily give a global

minimum of the sum of the distances between the inferred tetraploid ancestor and the outgroup, on one hand, and this same ancestor and its modern descendant, on the other.

During the most recent work with *Populus*, I realized that the heuristic was systematically biased so that the solution was too influenced by the tetraploid descendant and not enough by the outgroup. I attempted to remove this bias by permitting certain choices during the evaluation of path groups within the algorithm that are not consistent with ordinary optimal halving, in the hope that this would optimize guided halving. Though this worked with simulated data, I was unable to achieve any benefit for the real data, a loose end in this work that should eventually be taken up.

10.1.2 Phylogenomics

Incorporating an outgroup into the halving analysis contains the first seeds of a phylogenetic analysis, since there are "almost" three genomes being compared, the outgroup genome and the two intermingled genomes from the tetraploidization event. When there are two outgroups, this becomes phylogenetic in a more explicit way. This became apparent (early on) when I was actually confronting the problem of incorporating polyploids and diploids within the same phylogenomic analysis. This led to the task of setting up a general framework for decomposing a given phylogenomic problem into a set of various possible local configurations of doubled and unduplicated genomes, each of which requires a different strategy for integrating genomic distance, halving and rearrangement median algorithms.

As part of this strategy, I proposed two frameworks for analyzing two WGD descendants, one where the doubling precedes a speciation event and another where doubling occurs independently in both lineages initiated by a speciation event. I developed combi-

natorial algorithms permitting me to decide which of these options best explains the data.

10.1.3 Data quality

Combinatorial optimization algorithms for genome rearrangement generally require unambiguous gene order data where each gene is present once in each genome, and they are all accurately mapped. Real data, however, may not be so neat. I know this well from my previous work [237] on comparing genomes with partially ordered chromosomes [172, 235, 236] and on removing noise in comparative mapping [39, 239].

From the outset of my work, issues pertaining to data quality were encountered, especially with regards to the cereal data, which were resolved in principled in somewhat *ad hoc* way, and then with the angiosperm data. In the latter work, because of the size of the data sets, I felt it worthwhile and feasible to study some of these problems in a systematic way.

Homology sets. Can we use defective sets of homologs, i.e., which have only one copy in the duplicated genome or are missing the ortholog completely in the guide genome?

Singletons. Should we purge singletons from the data, i.e., sets of homologous markers who have no homologous adjacent markers in common in the either the duplicated genome or the outgroup?

Contigs. Can we use guide genomes that are not fully assembled, but are available only as sets of hundreds or thousands of contigs?

From the previous work [39, 239], we know that singletons generally add no useful information about evolutionary history. Including an additional singleton gene in the data tends to require the same number of extra steps as if it were moved to a random location, and does not confirm or invalidate whatever is inferred without it. In the angiosperm analysis I obtained this result in two ways, by calculating not only the genomic distance but also the *re-use* statistic.

Of interest is that the use of defective homology sets also seems to add noise to the solution, but careful analysis shows that this is due to a disproportionate number of singletons in these data sets. In fact, the defective homology sets do contain valuable information about rearrangement history.

Finally, I found a way of comparing genomes based on contig information only. This turns out to be surprisingly simple. Recently, my ideas about this problem have been discussed in an independent paper [141].

10.1.4 Complexity

In much of this thesis I have been more concerned with the running time of my heuristic algorithms on increasingly large data sets, and less with their theoretical worst-case behaviour. Nevertheless, I attempted to calculate the latter characteristic in many cases and, in Chapter 5, I derived the first complexity result for a guided genome halving problem. Inspired by this, Eric Tannier from Lyon asked me to collaborate further on this sort of problem. This project culminated in the large number of theoretical results in Chapter 9. Though the complexity of DCJ guided halving remains open, it is now clear that it is closely related to the DCJ median problem for multichromosomal genomes, which is also

open despite the misconception of many that it is solved. Both are likely NP-hard. We did, however, find a version of the guided halving problem which is polynomial, namely that pertaining to breakpoint distance and genomes unrestricted as to linearity or circularity of multiple chromosomes.

10.2 Ongoing and future work

10.2.1 Duplicate gene loss

Following whole genome duplication, the doubled genome undergoes a re-diploidization process. During this, some duplicate genes may be conserved as is, many will be permuted out of their original duplicated contexts through genome rearrangement, some may undergo neofunctionalization or subfunctionalization, and some may be effectively deleted, either through pseudogenization or through some physical mechanisms involving unequal recombination. I became interested in the spatial distribution of duplicate gene loss along the chromosome out of my concern for accurate reconstruction of the tetraploid ancestor. This led to a new way of assessing neighbourhood selection constraints on duplicate gene loss in Chapter 7, my first foray into the functional aspects of genome rearrangement, which has hitherto been a strictly structural enterprise.

I have by now confirmed the patterns found in *Populus* show up as well in *Saccharomyces*, though not in *Candida glabrata*, and I am also analyzing the *Paramecium* genome in this respect. In particular, I am investigating how the processes of gene loss are distributed throughout the genome and within a chromosome; detect if a duplicate gene loss is correlated to the retention or loss of nearby genes on the same chromosome; and compare

the gene loss selection patterns for different genomes. I hope to formulate and test hypotheses about the relative preponderance of pseudogenization and deletion phenomena, based on statistics on conserved "frame boundaries", where consecutive duplicates in two chromosomes are interrupted by one or more single genes in one or both of the chromosomes, an idea I proposed in Chapter 7.

10.2.2 The pathgroup approach

In the course of my work on guided halving, I developed the notion of a "pathgroup" as a way of simultaneously bringing to bear multiple sources of information at a decision point in a reconstruction algorithm. This proved to be a simple, effective and versatile tool. In particular it provides a powerful look-ahead capability for stepwise heuristics and a way of pruning search trees during exact searches of genome-like structures.

Moreover, the pathgroup formulation allows us to deal in a flexible way with defective homology sets, something of great importance for genome reconstruction using realistic data to full advantage.

I intend to introduce this structure into the study of the median problem and other phylogenetic tree inference problems. This could possibly include the "large" phylogeny problem, where not only the ancestral genomes must be reconstructed, but also the branching structure of the phylogeny.

10.2.3 Partial genome duplication

A recurrent counter-hypothesis to WGD that has been proposed for yeast, *Arabidopsis* and other organisms is that a limited number of chromosomes or large fragments of chromo-

somes doubled at various times during evolutionary history. This partial genome duplication process differs from WGD but shares some elements in common, in contrast to the independent duplication of individual genes. Though the genome halving approach cannot apply directly, it will provide a starting point for new methods to reconstruct evolutionary history. In particular, the pathgroups approach, which can withstand the abandonment of the "two copies of every gene" assumption, seems ideally suited for this enterprise.

Bibliography

[1] Adam Z, Sankoff D. 2008. The ABCs of MGR with DCJ. Evolutionary Bioinformatics, 4: 69–74.

[2] Adams KL, Percifield R, Wendel JF. 2004. Organ-specific silencing of duplicated genes in a newly synthesized allotetraploid. Genetics 168: 2217–2226.

[3] Adams KL, Wendel JF. 2005. Polyploidy and genome evolution in plants. Current Opinion Plant Biology 8: 135–141.

[4] Adams KL. 2007. Evolution of duplicated gene expression in polyploid and hybrid plants. Journal of Heredity 98: 136–141.

[5] Albert VA, Soltis DE, Carlsom JE, Farmerie WG, Wall PK *et al.* 2005. Floral gene resources from basal angiosperms for comparative genomics research. BMC Plant Biology 5: 5–16.

[6] Alekseyev MA, Pevzner PA. 2004. Genome halving problem revisited, In Lodaya, K., Mahajan, M., eds. Proceedings of FSTTCS 2004: Foundations of Software Technology and Theoretical Computer Science. Lecture Notes in Computer Science, 3328: 1–15. Heidelberg, Springer.

[7] Alekseyev M, Pevzner PA. 2008. Colored de Bruijn graphs and the genome halving problem. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 4: 98–107.

[8] Alfaro ME, Santini F, Brock CD. 2007 . Do reefs drive diversification in marine teleosts? Evidence from the pufferfish and their allies (Order Tetraodontiformes). Evolution 61: 2104–2126.

[9] APG II. 2003. An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants. Botanical Journal of the Linnean Society 141: 399–436.

[10] ARABIDOPSIS GENOME INITIATIVE. 2000. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. Nature 408: 796–815.

[11] Arroyo-García R, Ruiz-García L, Bolling L, Ocete R, López MA, et al. 2006. Multiple origins of cultivated grapevine (Vitis vinifera L. ssp. sativa) based on chloroplast DNA polymorphisms. Molecular Ecology 15: 3707–3714.

[12] Aury JM, Jaillon O, Duret L, Noel B, Jubin C, et al. 2006. Global trends of whole-genome duplications revealed by the ciliate Paramecium tetraurelia. Nature 444: 171–178.

[13] Bafna V, Pevzner P. 1996. Genome rearrangements and sorting by reversals. SIAM Journal of Computing 25: 272–289.

[14] Barrier M, Baldwin BG, Robichaux RH, Purugganan MD. 1999. Interspecific hybrid ancestry of a plant adaptive radiation: allopolyploidy of the Hawaiian silversword al-

liance (Asteracelea) inferred from floral homeotic gene duplications. Molecular Biology and Evolution 16: 1105–1113.

[15] Bell CD, Soltis DE, Soltis P. 2005. The age of the angiosperms: A molecular time scale without a clock. Evolution 59: 1245–1258.

[16] Bennett MD, Leitch IJ, Price HJ, Johnston JS. 2003. Comparisons with *Caenorhabditis* ($\tilde{1}$00 Mb) and Drosophila ($\tilde{1}$75 Mb) using flow cytometry show genome size in Arabidopsis to be $\tilde{1}$57 Mb and thus $\tilde{2}$5 % Larger than the Arabidopsis Genome Initiative Estimate of 125 Mb. Annals of Botany 91: 547–557.

[17] Bergeron A, Mixtacki J, Stoye J. 2006. A unifying view of genome rearrangements. In Bücher P, Moret BME eds. Algorithms in Bioinformatics. Proceedings of WABI 2006, Lecture Notes in Computer Science 4175: 163–173.

[18] Bergeron A, Mixtacki, J, Stoye J. 2008. On computing the breakpoint reuse rate in rearrangement scenarios. In Proceedings of RECOMB-CG, Lecture Notes in Computer Science 5267: 226–240.

[19] Bergeron A, Mixtacki J, Stoye J. 2008. HP distance via Double Cut and Join distance. In Combinatorial Pattern Matching, proceedings of CPM 2008, Volume 5029 of Lecture Notes in Computer Science.

[20] Berman P, Karpinski M. 1999. On some tighter inapproximability results. In Automata, Languages and Programming, Lecture Notes In Computer Science 1644:705.

[21] Bernt M, Merkle D, Middendorf M. 2008. Solving the Preserving Reversal Median Problem. IEEE/ACM Transactions on Computational Biology and Bioinformatics 5: 332–347.

[22] Billingsley P. 1995. Probability and measure, 3rd edition. New York: Wiley-Interscience. Patrick Billingsley.

[23] Blanc G, Barakat A, Guyot R, Cooke R, Delseny I. 2000. Extensive duplication and reshuffling in the arabidopsis genome. Plant Cell 12: 1093–1101.

[24] Blanc G, Hokamp K, Wolfe KH. 2003. A recent polyploidy superimposed on older large-scale duplications in the Arabidopsis genome. Genome Research 13: 137–144.

[25] Blanc G, wolfe KH. 2004. Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes. Plant Cell 16: 1667–1678.

[26] Bourque G, Pevzner P. 2002. Genome-scale evolution: Reconstructing gene orders in the ancestral species. Genome Research, 12:26–36.

[27] Bowers JE, Chapman BA, Rong J, Paterson AH. 2003. Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. Nature 422: 433–438.

[28] Bremer K. 2002. Gondwanan evolution of the grass alliance of families (Poales). Evolution 56: 1374–1387.

[29] Bryant D. 1998. The complexity of the breakpoint median problem. Technical Report CRM–2579, Centre de recherches mathématiques, Université de Montréal. Montreal, Canada.

[30] Buzgo M, Soltis PS, Kim S, Soltis DE. 2005 . The making of the flower. Biologist 52: 149–154.

[31] Byrne KP, Wolfe KH. 2005. The Yeast Gene Order Browser: combining curated homology and syntenic context reveals gene fate in polyploid species. Genome Research,15: 1456–1461.

[32] Byrnes JK, Morris GP, Li WH. 2006. Reorganization of adjacent gene relationships in yeast genomes by whole-genome duplication and gene deletion. Molecular Biology and Evolution 23: 1136–1143.

[33] Cannon SB, Sterck L, Rombauts S, Sato S, Cheung F, *et al.* 2006. Legume genome evolution viewed through the *Medicago truncatula* and *Lotus japonicus* genomes. Proceedings of the National Academy of Sciences 103: 14959–14964.

[34] Cantino P, Doyle J, Graham S, Judd W, Olmstead R, Soltis DE, Soltis PS, Donoghue M. 2007. Towards a phylogenetic nomenclature of Tracheophyta. Taxon 56: 822–846.

[35] Caprara A. 2003. The reversal median problem. INFORMS Journal on Computing, 15: 93–113.

[36] Carr GD. 1998. Chromosome evolution and speciation in Hawaiian flowering plants. In Stuessy TF and Ono M, eds. Evolution and speciation of island plants, 5–47. Cambridge University Press, Cambridge, UK.

[37] Clausen J, Keck DD, Hiesey WM. 1945. Experimental studies on the nature of species II. Plant evolution through amphiploidy and autopolyploidy, with examples from the Madiinae. Carnegie Institute of Washington, Washington, DC.

[38] Chen X, Zheng J, Fu Z, Nan P, Zhong Y, and T Jiang. 2005. SL: Assignement of orthologous genes via genome rearrangement. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2: 302–315.

[39] Choi V, Zheng C, Zhu Q, Sankoff D. 2007. Algorithms for the extraction of synteny blocks from comparative maps. in Giancarlo R, Hannenhalli S, eds. Proceedings of the WABI 2007 Workshop on Algorithms in Bioinformatics. Lecture Notes in Bioinformatics 4645: 277–288, Springer.

[40] Crepet WL, Feldman GD. 1991. The earliest remains of grasses in the fossil record. American Journal of Botany 78: 1010–1014.

[41] Cui L, Wall PK, Leebens-Mack JH, Lindsay BG, Soltis DE *et al.* 2006. Widespread genome duplications throughout the history of flowering plants. Genome Research 16: 738–749.

[42] Darlington CD. 1937. Recent advances in cytology. Second edition. P. Blakiston's, Philadelphia, PA.

[43] De Bodt S, Maere S, Van de Peer Y. 2005. Genome duplication and the origin of angiosperms. Trends Ecology Evolution 20: 591–597.

[44] Dietrich FS, Voegeli S, Brachat S, *et al.* 2004. The *Ashbya gossypii* genome as a tool for mapping the ancient *Saccharomyces cerevisiae* genome. Science 304: 304–307.

[45] Doyle JA. 2008. Integrating molecular phylogenetic and paleobotanical evidence on origin of the flower. International Journal of Plant Sciences 169: 816–843.

[46] Dujon B, Sherman D, Fischer G, *et al.* 2004. Genome evolution in yeasts. Nature 430: 35–44.

[47] El-Mabrouk N, Nadeau JH, Sankoff D.1998. Genome halving. In Farach-Colton M, ed. Combinatorial Pattern Matching, Ninth Annual Symposium. Lecture Notes in Computer Science 1448: 235–250. Heidelberg, Springer.

[48] El-Mabrouk N, Sankoff D. 1999. On the reconstruction of ancient doubled circular genomes using minimum reversals. In Asai K, Miyano S, Takagi T, eds. Genome Informatics 1999. Tokyo: Universal Academy Press 83–93.

[49] El-Mabrouk N, Sankoff D. 1999. Hybridization and genome rearrangement. In Crochemore M, Paterson M, eds. Combinatorial Pattern Matching. Tenth Annual Symposium. Lecture Notes in Computer Science 1645: 78–87.

[50] El-Mabrouk N, Bryant D, Sankoff D. 1999. Reconstructing the pre-doubling genome. In Istrail S, Pevzner P, Waterman M, eds. Proceedings of the Third Annual International Conference on Computational Molecular Biology (RECOMB 99). New York:ACM Press 154–163.

[51] El-Mabrouk N, Sankoff D. 2003. The reconstruction of doubled genomes. SIAM Journal on Computing 32: 754–792.

[52] Fertin G, Labarre A, Rusu I, Tannier E, vialette S. 2009. Combinatorics of Genome Rearrangements. MIT Press. in press.

[53] Friis EM, Pedersen KR, Crane PR. 2006 . Cretaceous angiosperm flowers: Innovation and evolution in plant reproduction.Paleogeography, Paleoclimatology, Paleoecology 232: 251–293.

[54] Frohlich MW, Chase MW. 2007. After a dozen years of progress the origin of angiosperms is still a great mystery. Nature 450: 1184–1189 .

[55] Gallardo MH, Bickham JW, Honeycutt RL, Ojeda RA, Kohler N. 1999. Discovery of tetraploidy in a mammal. Nature, 401:341.

[56] Gallardo MH, Kausel G, Jiménez A., Bacquet C., González C., Figueroa J., Kohler N, Ojeda R. 2004. Whole-genome duplications in South American desert rodents (*Octodontidae*). Biological Journal of the Linnean Society, 82: 443–451.

[57] Garey MR, Johnson DS. 1979. Computers and intractability. A guide to the theory of NP-completness.

[58] Gates RR. 1909. The stature and chromosomes of *Oenothera gigas* De Vries. Arch. f. Zellforsch 3: 525–552.

[59] Gaut BS, Morton BR, McCaig BC, Clegg MT. 1996. Substitution rate comparisons between grasses and palms: synonymous rate differences at the nuclear gene Adh parallel rate differences at the plastid gene rbcL. Proceedings of the National Academy of Sciences USA 93: 10274–10279.

[60] Gaut BS, Doebley JF. 1997. DNA sequence evidence for the segmental allotetraploid origin of maize. Proceedings of the National Academy of Sciences USA 94: 6809–6814.

[61] Gaut BS. 2001. Patterns of chromosomal duplication in maize and their implications for comparative maps of the grasses. Genome Research,11: 55–66.

[62] Goffeau A, Barrell BG, Bussey H *et al.* 1996. Life with 6000 genes. Science, 275: 1051–1052.

[63] Goldblatt P. 1980. Polyploidy in angiosperms: Monocotyledons. In Lewis WH, ed. Polyploidy: Biological relevance, 219–239. Plenum Press, New York, New York, USA.

[64] Gower JC. 1966. Some distance properties of latent root and vector methods used in multivariate analysis. Biometrika, 53: 325–328.

[65] Grant V. 1963. The Origin of Adaptations, New York, Columbia University Press.

[66] Grant V. 1981. Plant speciation. Second edition. Columbia University Press, New York, NY.

[67] GRAPPA(Genome Rearrangements Analysis under Parsimony and Other Phylogenetic Algorithms.) Available at http://www.cs.unm.edu/~moret/GRAPPA/. (Date last accessed May 2008)

[68] GRASS PHYLOGENY WORKING GROUP. 2001. Phylogeny and subfamilial classification of the grasses (Poaceae). Annals Missouri Botanical Garden 88: 373–457.

[69] Gregory TR, Mable BK. 2005. Polyploidy in animals. In Gregory TR, eds. The evolution of the genome 428–501. Elsevier Academic Press, San Diego, CA.

[70] Hannenhalli S, Pevzner P.1995. Transforming men into mice (polynomial algorithm for genomic distance problem). In Proceedings of the 36th Annual Symposium on Foundations of Computer Science FOCS. 581–592.

[71] Hannenhalli S, Pevzner P. 1999. Transforming Cabbage into Turnip: Polynomial Algorithm for Sorting Signed Permutations by Reversals. Journal of the ACM, 46:1–27.

[72] Harmon LJ, Weir JT, Brock CD, Glor RE, Challenger W. 2008. GEIGER: investigating evolutionary radiations. Bioinformatics 24: 129–131.

[73] Hughes AL. 1999. Phylogenies of developmentally important proteins do not support the hypothesis of two rounds of genome duplication early in vertebrate history. Journal of Molecular Evolution, 48: 565–576.

[74] Hughes NF. 1994. The Enigma of Angiosperm Origins. Cambridge University Press, Cambridge, UK.

[75] Husband BC. 2004. The role of triploids in the evolutionary dynamics of mixed-ploidy populations. Biological Journal Linnean Society 82: 537–546.

[76] The Arabidopsis Genome Initiative. 2000. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. Nature 408: 796–815.

[77] INTERNATIONAL RICE GENOME SEQUENCING PROJECT. 2005. The map-based sequence of the rice genome. Nature 436: 793–800.

[78] Irish V. 2006. Duplication, diversification, and comparative genetics of angiosperm MADS-box genes. Advances in Botanical Research 44: 127–159.

[79] Jackson S, Rounsley S, Purugganan M. 2006. Comparative sequencing of plant genomes: Choices to make. The Plant Cell 18: 1100–1104.

[80] Jaillon O, Aury JM, Brunet F *et al.* 2004. Genome duplication in the teleost fish Tetraodon nigroviridis reveals the early vertebrate proto-karyotype. Nature 431: 946–957.

[81] Jaillon O, Aury JM, Noel B, *et al.* 2007. The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. Nature 449: 463–467. http://www.genoscope.cns.fr/externe/English/Projets/Projet ML/data/annotation/

[82] Jaiswal P, Ni J, Yap I *et al.* 2006. Gramene: a bird's eye view of cereal genomes. Nucleic Acids Research, 34:D717–23. URL: http://www.gramene.org

[83] Jansen RK, Cai Z, Raubeson LA, Daniell H, DePamphilis CW, *et al.* 2007. Analysis of 81 genes from 64 plastid genomes resolves relationships in angiosperms and identifies genome-scale evolutionary patterns. Proceedings of the National Academy of Sciences 104: 19369–19374.

[84] Janssen T, Bremer K. 2004. The age of major monocot groups inferred from 800+ rbcL sequences. Botanical Journal of the Linnean Society. 146: 385–398.

[85] Jean G, Nikolski M: Genome rearrangements: a correct algorithm for optimal capping. Information Processing Letters 2007, 104: 14–20.

[86] Kellis M, Birren B, Lander E. 2004. Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae.* Nature, 428: 617–624.

[87] Kim S, Albert VA, Yoo MJ, Farris JS, Soltis PS, Soltis DE. 2004. Pre-angiosperm duplication of floral genes and regulatory tinkering at the base of angiosperms. American Journal of Botany 9: 2102–2118.

[88] Kim S, Koh J, Yoo MJ, Kong H, Hu Y, Ma H, Soltis PS, Soltis DE. 2005. Expression of floral MADS-box genes in basal angiosperms: Implications on evolution of floral regulators and the perianth. Plant Journal 43: 724–744.

[89] Koch MA, Kiefer M. 2005. Genome evolution among cruciferous plants: A lecture from the comparison of the genetic maps of three diploid species–*Capsella rubella, Arabidopsis lyrata subsp Petraea, and A. thaliana.* American Journal of Botany 92: 761–767.

[90] Kowalski S, Lan TH, Feldmann K, Paterson A. 1994. Comparative mapping of *Arabidopsis thaliana* and *Brassica oleracea* chromosomes reveals islands of conserved gene order. Genetics 138: 499–510.

[91] Kramer EM, Dorit RL, Irish VF. 1998. Molecular evolution of genes controlling petal and stamen development: Duplication and divergence within the *APETALA3* and *PISTILLATA* MADS-box gene lineages. Genetics 149: 765–783.

[92] Kramer EM, Jaramillo MA, Di Stilio VS. 2004. Patterns of gene duplication and functional evolution during the diversification of the *AGAMOUS* subfamily of MADS box genes in angiosperms. Genetics 166: 1011–1023.

[93] Kramer EM, Zimmer EA. 2006. Gene duplication and floral developmental genetics of basal eudicots. Advances in Botanical Research 44: 354–376.

[94] Kuittinen H, De Haan AA, Vogl C, Oikarinen S, Leppala J, *et al.* 2004. Comparing the linkage maps of the close relatives *Arabidopsis lyrata* and *A. thaliana.* Genetics 168: 1575–1584.

[95] Kurtzman CP, Robnett CJ. 2003. Phylogenetic relationships among yeasts of the "Saccharomyces complex" determined from multigene sequence analyses. FEMS Yeast Research, 3: 417–432.

[96] Kuwada Y. 1911. Meiosis in the pollen mother cells of *Zea mays* L. Botanical Magazine, Tokyo 25: 163–181.

[97] Lagercrantz U, Lydiate DJ. 1996. Comparative genome mapping in *Brassica.* Genetics 144: 1903–1910.

[98] Lan TH, Delmonte TA, Reischmann KP, Hyman J, Kowalski SP *et al.* 2000. An EST-enriched comparative map of *Brassica oleracea* and *Arabidopsis thaliana.* Genome Research 10: 776–788.

[99] Lavin M, Herendenn P, Wojciechowski MF. 2005. Evolutionary rates analysis of Leguminosae implicates a rapid diversification of lineages during the Tertiary. Systematic Biology 54: 575–594.

[100] Lawrence WJC. 1931. The secondary association of chromosomes. Cytologia 2: 342–384.

[101] LeComber SCL, Smith C. 2004. Polyplooidy in fishes: patterns and processes. Biological Journal of the Linnean Society 82: 431–442.

[102] Leebens-Mack JH, Raubeson LA, Cui L, Kuehl JV, Fourcade MH, *et al.* 2005. Identifying the basal angiosperm node in chloroplast genome phylogenies: Sampling one's way out of the Felsenstein zone. Molecular Biology and Evolution 22: 1948–1963.

[103] Leebens-Mack JH, Wall K, Duarte J, Zheng Z, OppenHeimer D, Depamphilis C. 2006. A genomics approach to the study of ancient polyploidy and floral developmental genetics. Advances in Botanical Research 44: 526–540.

[104] Leitch AR, Soltis DE, Soltis PS, Leitch IJ, Pires JC. 2004. Biological relevance of polyploidy: ecology to genomics. Biological Journal of the Linnean Society 82: 409–700.

[105] Leitch AR, Leitch IJ. 2008 . Genomic plasticity and the diversity of polyploid plants. Science 320: 481–483.

[106] Leitch IJ, Bennett MD. 2004. Genome downsizing in polyploid plants. Biological Journal of the Linnean Society 82: 651–663.

[107] Lenne R, Solnon C, Stutzle T, Tannier E, Birattari M. 2008. Reactive stochastic local search algorithms for the genomic median problem. In Proceedings of EvoCOP 2008, Volume 4972 of Lecture Notes in Computer Science 266–276.

[108] Levin DA. 1983. Polyploidy and novelty in flowering plants. American Naturalist 122: 1–25.

[109] Levin DA. 2000. The origin, expansion and demise of plant species. Oxford University Press, New York, NY.

[110] Levin DA. 2002. The role of chromosomal change in plant evolution. Oxford University Press, New York, NY.

[111] Lewis WH. 1980. Polyploidy in angiosperms: Dicotyledons. In Lewis WH, eds. Polyploidy: Biological relevance, 241–268. Plenum Press, New York.

[112] Lewis WH. 1980. Polyploidy in species populations. In Lewis WH, eds. Polyploidy biological relevance, 103–144. Plenum Press, New York.

[113] Li L, Stoeckert CJ Jr, Roos DS. 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Research 13:2178–2189.

[114] Lin Y, Lu C, Chang HY, Tang C. 2005. An efficient algorithm for sorting by block-interchange and its application to the evolution of vibrio species. Journal of Computational Biology 12: 102–112.

[115] Lin Y, Lu C, Liu Y, Tang C. 2006. SPRING: a tool for the analysis of genome rearrangement using reversals and block-interchanges, Nucleic Acids Research 34: W696–W699.

[116] Lindqvist C, Motley TJ, Jeffrey JJ, Albert VA. 2003. Cladogenesis and reticulation in the Hawaiian endemic mints (Lamiaceae). Cladistics 19: 480–495.

[117] Litt A, Irish VF. 2003. Duplication and diversification in the *APETALA1/FRUITFULL* floral homeotic gene lineage: implications for the evolution of floral development. Genetics 165: 821–833.

[118] Liu B, Wendel JF. 2003. Epigenetic phenomena and the evolution of plant allopolyploids. Molecular Phylogenetics and Evolution 29: 365–379.

[119] Löve A, Löve D. 1949. The geobotanical significance of polyploidy. I. Polyploidy and latitude. Portugaliae Acta Biologica (A) R. B. Goldschmidt special volumn: 273–352.

[120] Lovasz L, Plummer MD. 1986. Matching Theory. Elsevier Science Ltd.

[121] Lukens L, Quijada P, Udall J, Pires JC, Schranz ME, Osborn TC. 2004. Genome redundancy and plasticity within ancient and recent *Brassica* crop species. Biological Journal of the Linnean Society 82: 665–674.

[122] Lutz AM. 1907. A preliminary note on the chromosomes of *Oenothera lamarckiana* and one of its mutants, O. gigas. Science 26: 151–152.

[123] Lynch M, Force A. 2000. The probability of duplicate gene preservation by subfunctionalization. Genetics 154: 459–473.

[124] Lynch M, Conery JS. 2000. The evolutionary fate and consequences of duplicate genes. Science 290: 1151–1155.

[125] Mable BK. 2004. Why polyploidy is rarer in animals that in plants: Myths and mechanisms. Biological Journal of the Linnean Society 82: 453–466.

[126] Magallón S, Sanderson MJ. 2001. Absolute diversification rates in angiosperm clades. Evolution 55: 1762–1780.

[127] Magallón S, Sanderson MJ. 2005. Angiosperm divergence times: the effect of genes, codon positions, and time constraints. International Journal of Organic Evolution 59: 1653–1670.

[128] Malcomber ST, Kellogg EA. 2005. *SEPALLATA* gene diversification: Brave new whorls. Trends Plant Science 10: 427–435.

[129] Marchant AD, Briggs BG. 2007. Ecdeiocoleaceae and Joinvilleaceae, sisters of Poaceae (Poales): Evidence from *rbc*L and *mat*K data. Telopea 11: 437–450.

[130] Masterson J. 1994. Stomatal size in fossil plants: Evidence for polyploidy in majority of angiosperms. Science 264: 421–423.

[131] Mazowita M, Haque L, Sankoff D. 2006. Stability of rearrangement measures in the comparison of genome sequences. Journal of Computational Biology13: 554–566.

[132] Mcelwain JC, Punyasena SW. 2007. Mass extinction events and the plant fossil record. Trends in Ecology and Evolution 22: 548–557.

[133] Mcgrath JM, Jancso MM, Pichersky E. 1993. Duplicate sequences with a similarity to expressed genes in the genome of *Arabidopsis thaliana*. Theoretical and Applied Genetics 86: 880–888.

[134] McLysaght A, Hokamp K, Wolfe KH. 2002. Extensive genomic duplication during early chordate evolution. Nature Genetics, 31: 200–204.

[135] Ming R, Hou S, Feng Y, Yu Q, Dionne-Laporte A, *et al.* 2008. The draft genome of the transgenic tropical fruit tree papaya (*Carica papaya* Linnaeus). Nature 452: 991–996. http://asgpb.mhpcc.hawaii.edu

[136] Mixtacki J. 2008. Genome Halving under DCJ revisited. In Proceedings of COCOON 2008, Lecture Notes in Computer Science.

[137] Moore G, Devos KM, Wang Z, Gale MD. 1995. Cereal genome evolutio–Grasses, line up and form a circle. Current Biology 5:737–739.

[138] Moore MJ, Bell CD, Soltis PS, Soltis DE. 2007. Using plastid genomic-scale data to resolve enigmatic relationships among basal angiosperms. Proceedings of the National Academy of Sciences USA 104:19363–19368.

[139] Moret B, Tang J, Warnow T, 2005. Reconstructing phylogenies from gene-content and gene-order data. In: Gascuel, O. (Ed.), Mathematics of Evolution and Phylogeny. Oxford University Press: 321–352.

[140] Muffato M, Crollius HR. 2008. Paleogenomics in vertebrates, or the recovery of lost genomes from the mist of time. Bioessays 30: 122–134.

[141] Muñoz A, Sankoff D. 2009. Rearrangement phylogeny of genomes in contig form. Accept by International Symposium on Bioinformatics Research and Applications 2009.

[142] Müntzing A. 1936. The evolutionary significance of autopolyploidy. Hereditas 21: 263–378.

[143] Nadeau JH, Sankoff D. 1997. Comparable rates of gene loss and functional divergence after genome duplications early in vertebrate evolution. Genetics, 147: 1259–1266.

[144] Ohno S, Wolf U, Atkin NB 1968. Evolution from fish to mammals by gene duplication, Hereditas 59: 169–187.

[145] Olmstead RG, Bohs L, Migid HA, Santiago-Valentin E, Garcia VF, Collier SM. 2009. A molecular phylogeny of the Solanaceae. Taxon, In press.

[146] Osborn TC, Pires JC, Birchler JA, Auger DL, Chen ZJ *et al.* 2003. Understanding mechanisms of novel gene expression in polyploids. Trends in Genetics 19: 141–147.

[147] Otto SP, Whitton J. 2000. Polyploid incidence and evolution. Annual Review of Genetics 34: 401–437.

[148] Ozery-Flato M, Shamir R. 2003. Two notes on genome rearrangement. Journal of Bioinformatics and Computational Biology 1: 71–94.

[149] Ohlebusch E, Abouelhoda MI, Hockel K. 2007. A linear time algorithm for the inversion median problem in circular bacterial genomes. Journal of Discrete Algorithms 5: 637–646.

[150] Paterson AH, Bowers JE, Burow M, Draye X, Elsik C, *et al.* 2000. Comparative genomics of plant chromosomes. Plant Cell 12: 1523–1539.

[151] Paterson AH, Bowers JE, Chapman BA. 2004. Ancient polyploidization predating divergence of the cereals, and its consequences for comparative genomics. Proceedings of the National Academy of Sciences 101: 9903–9908.

[152] Paterson AH, Chapman BA, Kissinger JC, Bowers JE, Feltus FA, and Estill JC. 2006. Many gene and domain families have convergent fates following independent whole-genome duplication events in *Arabidopsis, Oryza, Saccharomyces* and *Tetraodon.* Trends in Genetics 22: 597–602 .

[153] Pe'er I, Shamir R. 1998. The median problems for breakpoints are NP- complete. Electronic Colloquium on Computational Complexity Technical Report 98-071, Available at http://www.eccc.uni-trier.de/eccc. (Date last accessed May 2008).

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

[155] Pevzner P, Tesler G. 2003. Transforming men into mice: the Nadeau-Taylor chromosomal breakage model revisited. In Proceedings of the seventh annual international conference on Research in computational molecular biology RECOMB 2003: 247–256.

[156] Pfeil BE, Schlueter JA, Shoemaker RC, Doyle JJ. 2005. Placing paleopolyploidy in relation to taxon divergence: A phylogenetic analysis in legumes using 39 gene families. Systematic Biology 54: 441–454.

[157] Postlethwait JH, Yan YL, Gates MA, Horne S, Amores A *et al.* 1998. Vertebrate genome evolution and the zebrafish gene map. Nature Genetics 18: 345–349.

[158] Poyatos JF, Hurst LD. 2007. The determinants of gene order conservation in yeasts. Genome Biology 8: R233.

[159] Pryer KM, Schneider H, Zimmer EA, Banks JA. 2002. Deciding among green plants for whole genome studies. Trends in Plant Science 7: 550–554.

[160] Quiros CF, Grellet F, Sadowski J, Suzuki T, Li G, Wroblewski T. 2001. *Arabidopsis* and *Brassica* comparative genomics: Sequence, structure and gene content in the *ABI1-Rps2-Ck1* chromosomal segment and related regions. Genetics 157: 1321–1330.

[161] R Development Core Team. 2007. R: A language and environment for statistical computing. R Foundation for Statistical Computing. Available at http://www.R-project.org. (Date last accessed May 2008).

[162] Ramsey JR, Schemske DW. 1998. Pathways, mechanisms, and rates of polyploid formation in flowering plants. Annual Review of Ecology and Systematics 29: 467–501

.

[163] Ramsey JR, Schemske DW. 2002. Neopolyploidy in flowering plants. Annual Review of Ecology and Systematics 33: 589–639.

[164] Rapp RA, Wendel JF. 2005. Epigenetics and plant evolution. New Phytologist 168: 81–91.

[165] Rieseberg LH, Willis JH. 2007. Plant speciation. Science 317: 910–914.

[166] Saarela JM, Rai HS, Doyle JA, Endress PK, Mathews S *et al.* 2007. Hydatellaceae identified as a new branch near the base of the angiosperm phylogenetic tree. Nature 446: 312–315.

[167] Sampedro J, Lee Y, Carey RE, DePamphilis C, Cosgrove DJ. 2005. Use of genomic history to improve phylogeny and understanding of births and deaths in a gene family. Plant Journal 44: 409–19.

[168] Sanderson MJ, Doyle JA. 2001. Sources of error and confidence intervals in estimating the age of angiosperms from *rbc*L and 18S rDNA data. American Journal of Botany 88: 1499–1516.

[169] Sankoff D, Blanchette M. 1997. The median problem for breakpoints in comparative genomics. In Jiang T, Lee DT, eds. Proceedings of the Third International Computing and Combinatorics Conference COCOON 1997. Lecture Notes in Computer Science 1276:251–263.

[170] Sankoff D, and Blanchette M. 1998. Multiple genome rearrangement and breakpoint phylogeny. Journal of Computational Biology 5: 555–570.

[171] Sankoff D, El-Mabrouk N. 2000. Duplication, rearrangement and reconciliation. Comparative Genomes (Sankoff, D. Nadeau, J. eds), 537–550.

[172] Sankoff D, Zheng C Lenert A. 2005. Reversals of fortune. In McLysaght, A. and Huson, D. (ed), RECOMB 2005 Ws on Comparative Genomics, LNBI 3678. Berlin, Heidelberg:Springer Verlag,LNBI 3678: 131-141.

[173] Sankoff D. 2006. The signal in the genomes. PLoS Computional Biology 2: e35.

[174] Sankoff D, Zheng C, Zhu Q. 2007. Polyploids, genome halving and phylogeny. Bioinformatics, 23, i433–i439.

[175] Sankoff D, Zheng C, Wall PK, dePamphilis C, Leebens-Mack J, Albert V. 2008. Towards improved reconstruction of ancestral gene order in angiosperm phylogeny. Submitted to Journal of Computational Biology.

[176] Sankoff D, Zheng C, Wall PK, dePamphilis C, Leebens-Mack J, Albert V. 2008. Internal validation of ancestral gene order reconstruction in angiosperm phylogeny. Proceedings of RECOMB-CG 2008, Lecture Notes in Computer Science 5267:252–264, Springer.

[177] Scannell DR, Frank AC, Conant GC, Byrne KP, Woolfit M, Wolfe KH. 2007. Independent sorting-out of thousands of duplicated gene pairs in two yeast species descended from a whole-genome duplication. Proceedings of the National Academy of Sciences USA 104(20): 8397–402.

[178] Schlueter JA, Dixon P, Granger C, Grant D, Clark L, Doyle JJ, Schoenmaker RC. 2004. Mining EST databases to resolve evolutionary events in major crop species. Genome 47: 868–876.

[179] Schranz EM, Mitchell-olds T. 2006. Independent ancient polyploidy events in the sister families Brassicaceae and Cleomaceae. The Plant Cell 18: 1152–1165.

[180] Seoighe C, Wolfe KH. 1998. Extent of genomic rearrangement after genome duplication in yeast. Proceedings of the National Academy of Sciences USA 95:4447–4452.

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

[182] Siepel A, Moret BME. 2001. Finding an optimal inversion median: experimental results. In Gascuel O, Moret BME eds. Algorithms in Bioinformatics. Proceedings of WABI 2001, Lecture Notes in Computer Science, 2149: 189–203.

[183] Simillion C, Vandepoele K, Van Montagu MCE, Zabeau M, Van De Peer Y. 2002. The hidden duplication past of *Arabidopsis thaliana*. Proceedings of the National Academy of Sciences USA 99: 13627–13632.

[184] Sinha AU, Meller J. 2008. Sensitivity analysis for reversal distance and breakpoint reuse in genome rearrangements. Pacific Symposium on Biocomputing: 37–48.

[185] Smith SA, Donoghue MJ. 2008. Rates of molecular evolution are linked to life history in owering plants. Science 322: 86–89.

[186] Soltis DE, Soltis PS. 1990. Isozyme evidence for ancient polyploidy in primitive angiosperms. Systematic Botany 15: 328–337.

[187] Soltis DE, Soltis PS. 1999. Polyploidy: Origins of species and genome evolution. Trends in Ecology & Evolution 14: 348–352 .

[188] Soltis DE, Soltis PS, Chase MW, Mort ME, Albach DC, *et al.* 2000. Angiosperm phylogeny inferred from a combined data set of 18S rDNA, *rbcL*, and *atpB* sequences. Botanical Journal of the Linnean Society 133: 381–461.

[189] Soltis DE, Soltis PS, Tate JA. 2003. Advances in the study of polyploidy since Plant Speciation. New Phytologist 161: 173–191.

[190] Soltis DE, Soltis PS, Chase MW, Endress PK. 2005. Phylogeny, Evolution, and Classification of Flowering Plants. Sinauer Associates, Sunderland, MA.

[191] Soltis DE, Soltis PS, Schemske DW, Hancock JF, Thompson JN *et al.* 2007. Autopolyploidy in angiosperms: Have we grossly underestimated the number of species? Taxon 56: 13–30.

[192] Soltis DE, Albert VA, Leebens-Mack JH, Palmer J, Wing R, et al. 2008. The *Amborella* Genome Initiative: A genome for understanding the evolution of angiosperms. Genome Biology 9: 402.

[193] Soltis DE, Bell CD, Kim S, Soltis PS. 2008. The origin and early evolution of the angiosperms. Annals of the New York Academy of Sciences 1133:3–25.

[194] Soltis DE, Leebens-Mack J, Bella CD, Paterson A, Albert VA, Zheng C, D. Sankoff D, Soltis PS. 2009. Polyploidy and angiosperm diversication. American Journal of Botany 96: 336–348.

[195] Soltis PS, Soltis DE, Chase MW. 1999. Angiosperm phylogeny inferred from multiple genes: A research tool for comparative biology. Nature 402: 402–404.

[196] Soltis PS, Soltis DE. 2000. The role of genetic and genomic attributes in the success of polyploids. Proceedings of the National Academy of Sciences USA 97: 7051–7057.

[197] Soltis PS, Soltis DE, Kim S, Chanderbali A, Buzgo M. 2006. Expression of floral regulators in basal angiosperms and the origin and evolution of ABC-function. Advances in Botanical Research 44: 323–347.

[198] Spangler R, Zaitchik B, Russo E, Kellogg E. 1999. Andropogoneae evolution and generic limits in *Sorghum* (Poaceae) using *ndhF* sequences. Systematic Botany 24: 267–281.

[199] Stebbins GL Jr. 1940. The significance of polyploidy in plant evolution. American Naturalist 74: 54–66.

[200] Stebbins GL. 1947. Types of polyploids: Their classification and significance. Advances Genetics 1: 403–429.

[201] Stebbins GL. 1950. Variation and evolution in plants. Columbia University Press, New York, NY.

[202] Stebbins GL. 1971. Chromosomal evolution in higher plants. Addison-Wesley, London.

[203] Stebbins GL. 1985. Polyploidy, hybridization, and the invasion of new habitats. Annals Missouri Botanical Garden 72: 824–832.

[204] Stevens PF. 2007. Angiosperm Phylogeny Website. Version 8, June 2007. http://www.mobot.org/MOBOT/research/APweb/.

[205] Tang J, Moret B. 2003. Phylogenetic Reconstruction from Gene-Rearrangement Data with Unequal Gene Content. Lecture Notes in Computer Science,Volume 2748: 37–46.

[206] Tang H, Wang X, Bowers JE, Ming R, Alam M, Paterson AH. 2008. Unraveling ancient hexaploidy through multiply aligned angiosperm gene maps. Genome Research [Epub ahead of print] PMID: 18832442

[207] Tang H, Bowers JE, Wang X, Ming R, Alam M, Paterson AH. 2008. Synteny and collinearity in plant genomes. Science 320: 486–488.

[208] Tannier E, Zheng C, Sankoff D. 2008. Multichromosomal median and halving problems under different genomic distances. In Crandall KA, Lagergren J, eds. Workshop on Algorithms in Bioinformatics (WABI 2008). Lecture Notes in Bioinformatics 5251: 1–13.

[209] Tate JA, Soltis DE, Soltis PS. 2005. Polyploidy in plants. In Gregory TR, eds. The evolution of the genome. Elsevier Academic Press, San Diego, CA. 371–426.

[210] Tesler G. 2002. Efficient algorithms for multichromosomal genome rearrangements.Journal of Computer and System Sciences 65: 587–609.

[211] Thompson JN, Cunningham BM, Segraves KA, Althoff DM, Wagner D. 1997. Plant polyploidy and insect/plant interactions. American Naturalist 150: 730–743.

[212] Thompson JN, Nuismer SL, Merg K. 2004. Plant polyploidy and the evolutionary ecology of plant/animal interactions. Biological Journal Linnean Society 82: 503–510.

[213] Tuskan GA, Difazio S, Jansson S, Bohlmann J, Grigoriev I,*et al.* 2006. The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). Science 313: 1596–1604. http://genome.jgi-psf.org/Poptr1/Poptr1.download .html

[214] van Hoek MJ, Hogeweg P. 2007. The role of mutational dynamics in genome shrinkage. Molecular Biology and Evolution 24: 2485–2494.

[215] VandePeer Y, Meyer A. 2005. Large-scale gene and ancient genome duplications. In Gregory TR, eds. The evolution of the genome. Elsevier Academic Press, San Diego, CA. 329–370.

[216] Velasco R, Zharkikh A, Troggio M, Cartwright DA, Cestaro A, *et al.* 2007. A high quality draft consensus sequence of the genome of a heterozygous grapevine variety. PLoS ONE 2: e1326.

[217] Veron AS, Kaufmann K, Bornberg-Bauer E. 2007. Evidence of interaction network evolution by whole-genome duplications: A case study in MADS-box proteins. Molecular Biology and Evolution 24: 670–678.

[218] Vision TJ, Brown DG, Tanksley SD. 2000. The origins of genomic duplications in *Arabidopsis*. Science 290: 2114–2117.

[219] Warren R, Sankoff D. 2008. Genome halving with double cut and join. In Proceedings of the 6th Asia-Pacific Bioinformatics Conference, Advances in Bioinformatics and Computational Biology 6: 231–240.

[220] Watterson G, Ewens W, Hall T, Morgan A. 1982. The chromosome inversion problem. Journal of Theoretical Biology 1982.

[221] Wendel JF. 2000. Genome evolution in polyploids. Plant Molecular Biology 42: 225–249.

[222] Wendel JF, Doyle JJ. 2005. Polyploidy and evolution in plants. In Henry R, eds. Diversity and Evolution in Plants 97–117, CABI Publishing, Oxon, UK.

[223] Werth CR, Windham MD. 1991. A model for divergent, allopatric speciation of polyploid Pteridophytes resulting from silencing of duplicate-gene expression. American Naturalist 137: 515.

[224] Wikström N, Savolainen V, Chase MW. 2001. Evolution of the angiosperms: Calibrating the family tree. Proceedings of the Royal Society of London, B, Biological Sciences 268: 2211–2220.

[225] Wilson WA, Harrington SE, Woodman WL *et al.* 1999. Inferences on the genome structure of progenitor maize through comparative analysis of rice, maize and the domesticated panicoids. Genetics,153: 453–473.

[226] Wojciechowski MF, Lavin M, Sanderson MJ. 2004. A phylogeny of legumes (Leguminosae) based on analysis of the plastid *matK* gene resolves many well-supported subclades within the family. American Journal Botany 91: 1846–1862.

[227] Wolfe KH. Shields DC. 1997. Molecular evidence for an ancient duplication of the entire yeast genome. Nature, 387, 708–713.

[228] Woolfe A, Elgar G. 2007. Comparative genomics using *Fugu* reveals insights into regulatory subfunctionalizaation. Genome Biology 8: R53.

[229] Xu RH, Kim J, Taira M, Lin JJ, Zhang CH *et al.* 1997. Differential regulation of neurogenesis by the two Xenopus GATA-1 genes. Molecular and Cellular Biology, 17:436–443.

[230] Xu W, Sankoff D. 2008. Decompositions of Multiple Breakpoint Graphs and Rapid Exact Solutions to the Median Problem. In Algorithms in Bioinformatics, proceedings of WABI, Lecture Notes in Bioinformatics.

[231] Yancopoulos S, Attie O, Friedberg R. 2005. Efficient sorting of genomic permutations by translocation, inversion and block interchange. Bioinformatics, 21: 3340–3346

[232] Yogeeswaran K, Frary A, York TL, Amenta A, Lesser AH *et al.* 2005. Comparative genome analyses of *Arabidopsis* spp.: Inferring chromosomal rearrangement events in the evolutionary history of *A. thaliana.* Genome Research 15: 505–515.

[233] Yu J, Wang J, Lin W, Li S, Li H *et al.* 2005 . The genomes of *Oryza sativa*: A history of duplications. PLoS Biology 3: e38, doi:10.1371/journal.pbio.0030038.

[234] Zahn LM, Kong H, Leebens-Mack JH, Kim S, Soltis PS, *et al.* 2005. The evolution of the *SEPALLATA* subfamily of MADS-box genes: a pre-angiosperm origin with multiple duplications throughout angiosperm history. *Genetics* 169: 2209–2223.

[235] Zheng, C., Lenert, A. and Sankoff, D. 2005. Reversal distance for partially ordered genomes. Bioinformatics 21, pp. i502-i508.

[236] Zheng, C. and Sankoff D. 2005. Genome Rearrangements with partially ordered chromosomes. In L.Wang (ed.) Proceedings of COCOON 2005, LNCS, Springer, pp. 52-62.

[237] Zheng, C. 2006. Genome Rearrangement Algorithms Applied to Comparative Maps. Master thesis. University of Ottawa.

[238] Zheng C, Zhu Q, Sankoff D. 2006. Genome halving with an outgroup. Evolutionary Bioinformatics 2: 319–326.

[239] Zheng C, Zhu Q, Sankoff D. 2007. Removing noise and ambiguities from comparative maps in rearrangement analysis. IEEE/ACM Transactions in Computational Biology and Bioinformatics 4: 515–522.

[240] Zheng C, Zhu Q, Sankoff D. 2007. Parts of the problem of polyploids in rearrangement phylogeny. in Tesler G, Durand D, eds. RECOMB 2007 Workshop on Comparative Genomics. LNCS 4751:162–176, Springer.

[241] Zheng C, Zhu Q, Adam Z, Sankoff D. 2008. Guided genome halving: hardness, heuristics and the history of the Hemiascomycetes. Bioinformatics 24: i96–i104.

[242] Zheng C, Zhu Q, Sankoff D. 2008. Descendants of whole genome duplication within gene order phylogeny. Journal of Computational Biology 15:947–964.

[243] Zheng C, Wall PK, Leebens-Mack J, dePamphilis C, Albert VA, Sankoff D. 2008. The effect of massive gene loss following whole genome duplication on the algorithmic reconstruction of the ancestral Populus diploid. Proceedings of the Conference on Computational Systems Bioinformatics 2008, 261–271.

[244] Zheng C, Wall PK, Leebens-Mack J, dePamphilis C, Albert VA, Sankoff D. 2009. Gene loss under neighbourhood selection following whole genome duplication and the

reconstruction of the ancestral populus genome. Journal of Bioinformatics and Computational Biology, in press.