Computational Biology

Cedric Chauve
Nadia El-Mabrouk
Eric Tannier *Editors*

Models and Algorithms for Genome Evolution

ISCB INTERNATIONAL SOCIETY FOR COMPUTATIONAL BIOLOGY

Springer

The *Computational Biology* series publishes the very latest, high-quality research devoted to specific issues in computer-assisted analysis of biological data. The main emphasis is on current scientific developments and innovative techniques in computational biology (bioinformatics), bringing to light methods from mathematics, statistics and computer science that directly address biological problems currently under investigation.

The series offers publications that present the state-of-the-art regarding the problems in question; show computational biology/bioinformatics methods at work; and finally discuss anticipated demands regarding developments in future methodology. Titles can range from focused monographs, to undergraduate and graduate textbooks, and professional text/reference works.

Author guidelines: springer.com > Authors > Author Guidelines

For further volumes:
www.springer.com/series/5769

Cedric Chauve • Nadia El-Mabrouk • Eric Tannier
Editors

Models and Algorithms for Genome Evolution

Springer

Computational Biology

The *Computational Biology* series publishes the very latest, high-quality research devoted to specific issues in computer-assisted analysis of biological data. The main emphasis is on current scientific developments and innovative techniques in computational biology (bioinformatics), bringing to light methods from mathematics, statistics and computer science that directly address biological problems currently under investigation.

The series offers publications that present the state-of-the-art regarding the problems in question; show computational biology/bioinformatics methods at work; and finally discuss anticipated demands regarding developments in future methodology. Titles can range from focused monographs, to undergraduate and graduate textbooks, and professional text/reference works.

Author guidelines: springer.com > Authors > Author Guidelines

For further volumes:
www.springer.com/series/5769

Cedric Chauve • Nadia El-Mabrouk • Eric Tannier
Editors

Models and Algorithms for Genome Evolution

Springer

Chapter 11
Fractionation, Rearrangement, Consolidation, Reconstruction

David Sankoff and Chunfang Zheng

Abstract The reconstruction of ancestral gene orders based on models of chromosomal rearrangement mechanisms is complicated when some of the input genomes have undergone whole genome duplications followed by fractionation, the massive loss of some or most of the duplicate genes. We describe a reconstruction protocol that uses maximum weight matching in two phases to overcome the fragmented nature of results based on gene adjacency only. We review consolidation methods for recovering synteny patterns from fractionated genomes, and show how to integrate these into the reconstruction protocol. The procedure is applied to reconstruct the common ancestral gene order of grape and poplar. Simulation of the evolution of comparable genomes reveals the narrow ranges within which the rearrangement and fractionation parameters must be set in order to emulate statistical attributes of the extant genomes.

11.1 Introduction

All methods of reconstructing the details of ancestral gene order from a number of extant genomes are based on common gene adjacencies in these genomes, e.g., [1–4], though they all build on these fundamental data in different ways. As evolution progresses *rearrangement* events, notably inversion and reciprocal translocation, successively disrupt gene adjacencies in individual genomes. In addition, more local events such as gene transposition from one site to another, gene deletion and gene duplication also disrupt some adjacencies and establish others. To the extent that many common adjacencies remain unperturbed by these processes in two or more of the extant genomes, they may contain enough evolutionary signal to allow reconstruction of significant portions of the ancestral order. As evolution continues, rearrangement can eventually degrade this signal so that only relatively short fragments—"contigs"—of the ancestral order can be inferred. Reconstruction methods need to transcend their dependence on gene adjacencies if longer range gene orders are required.

D. Sankoff (✉) · C. Zheng
University of Ottawa, Ottawa, ON, Canada
e-mail: sankoff@uottawa.ca

C. Chauve et al. (eds.), *Models and Algorithms for Genome Evolution*,
Computational Biology 19, DOI 10.1007/978-1-4471-5298-9_11,

A complication arises if one or more of the extant genomes derive from whole genome duplication (WGD) events that occurred since their common ancestor. The duplication itself does not add any new adjacencies or remove any; the pre-existing adjacencies simply continue, but with multiplicity two. What complicates things after WGD is duplicate gene loss on a massive scale, deleting one or the other, but not both, of most duplicate gene pairs, a process called *fractionation* [5]. The loss of an individual gene y from context $\ldots xyz\ldots$ will generally destroy two adjacencies xy and yz and create a new one xz. This is true whether the loss of the gene is a physical deletion of part of the chromosome or by pseudogenization. Even if xy and yz still exist in the homeologous region of the genome, the adjacency xz is an innovation.

WGD and fractionation are particularly prevalent in flowering plants [6], where the slow (tens or hundreds of millions of years) cycle of the two processes also involves the constant excision of excess non-coding DNA, a characteristic of genome dynamics that distinguishes these organisms from other evolutionary domains, such as the mammals.

It is misleading to compare fractionated genomes with non-WGD relatives in terms of rearrangements only, because these yield systematically exaggerated results: the algorithms are forced to account for the missing and the new adjacencies as if they were breakpoints of (non-existent) inversions and translocations. And although there have long been methods for incorporating gene loss into genome rearrangement algorithms [7], these are not designed for the specific scenario of WGD followed by fractionation.

In this paper, we first detail our *scaffolding* approach [4] to overcoming the "short contigs" limitations of adjacency-based reconstruction (Sect. 11.2). After an explanation of the notion of *excess adjacencies* in Sect. 11.3, we briefly review WGD and fractionation (Sect. 11.4). In Sect. 11.5 we then present an improved *consolidation interval* strategy [8, 9] for accounting for fractionation in a descendant of a WGD event. We can then reconstruct an ancestral gene order for the pre-WGD genome from two or more of its direct descendants, where the genes in these descendant genomes are replaced by consolidated intervals. Finally, the genes inside these intervals are sorted.

We illustrate in Sect. 11.6 using the two plant genomes, from poplar (a WGD descendant) and grapevine (no recent WGD history), to reconstruct their common "rosid" ancestor, with and without taking into account fractionation. We find that the consolidation step removes virtually all the artifactual rearrangements inferred when fractionation is ignored.

From this reconstruction, we can calculate the total amount of rearrangement from the ancestor to the two extant genomes, how much this would be inflated by ignoring fractionation history, the distribution of sizes of consolidation intervals, and the fractionation bias—to what extent are genes deleted in an asymmetric manner from the two copies of the genome emerging from WGD.

In Sect. 11.7, we set up a simulation of the gene order evolution of poplar and grapevine from the ancestral rosid, with the same numbers of genes and chromosomes, and the same number of single-copy genes (= number of gene losses) in the simulated poplar genome. We experimented with parameters reflecting the numbers of rearrangement operations, what proportion of these are short inversions, how many genes are deleted at a time, and how probable a deletion is to affect one or the other of the original two copies of the poplar genome. We determined the unique set of evolutionary parameter values producing the observed values in our analysis of the real plant genomes.

11.2 Reconstruction

Our reconstruction method requires preprocessing annotated genomic data by a syntenic block detection program such as SynMap in the CoGe platform [10, 11] to identify likely orthologous genes in all pairs of the genomes under study, as well as paralogs in self-comparison of each descendant of recent WGD. We then process the combined set of all these orthologies and paralogies with the OMG! procedure [12] to produce homology sets containing at most N paralogous versions of each gene in each $2N$-ploid, including at most one gene in each diploid. We also impose some more or less stringent condition such as at least two genes from different genomes in each set. These homology sets represent candidate genes for the reconstructed ancestral genome. The use of stringent criteria in SynMap and OMG! ensure that each set can be mirrored by only one gene in the eventual reconstruction. Though this lends confidence to the reconstruction of the particular gene and its position in the gene order, it does exclude the possibility of assigning additional genes, even in a tentative way.

Once we have the set of relevant genes and all the homology relations we reconstruct the ancestral order using Maximum Weight Matching (MWM) [13] at two levels. First we identify all the gene adjacencies (considering only the genes within the data set as constructed) in all the genomes and subgenomes, each homology set determining two vertices of a graph G_1, corresponding to the 5' and 3' ends of the genes involved. We weight each adjacency—an edge in G_1—according to how many times homologs of the two genes involved are adjacent, with that particular 5'–3' orientation, in the data, possibly taking into account phylogenetic or data quality considerations, depending on the particular biological problem being analyzed. The MWM then chooses an optimal subset of adjacencies. This gives a set of ancestral "contigs". A small number of these may be circular; we linearize each of these by discarding their lowest weight adjacency—this has a minuscule effect on the total weight of the matching.

For the second application of MWM, we use the contigs as vertices in a graph G_2. Each contig has a mean position (as measured in gene order position) on a chromosome in one or more of the input genomes or subgenomes. These positions order the contigs on chromosomes. The few ambiguous contigs, i.e., containing large proportions of genes originating in two or more chromosomes in the same genome or subgenome are discarded. In addition, to ensure a level of syntenic robustness, if a

contig does not have a minimum number of genes in at least one genome, we discard it. Thresholds are set so that losing these small contigs plus the ambiguous ones satisfies a trade-off between accuracy and gene inclusiveness.

Two successive contigs on a chromosome are considered adjacent, and are joined by an edge in G_2 for the purposes of the second MWM. The orientation of a contig on a chromosome is determined by whether the genes it contains are in largely increasing or decreasing gene order on the subgenome in question. The weights may be the same as in the first MWM, or may be different. The output from this algorithm is a set of "scaffolds", namely a series of contigs alternating with gaps, each corresponding to a chromosome or a fragment of a chromosome in the ancestral genome.

Linearizing circular scaffolds turns out to be a quantitatively more important problem than with contigs, Nevertheless, we have found that shifting the adjacency criterion in mid-analysis from gene adjacency to contig proximity is an effective way of transcending the short contig limitation of ancestral reconstruction [4].

Though the principle of MWM has been used for ancestral genome reconstruction in a variety of theoretical contexts [14–16], it is also well suited to practical problem of scaling up from the gene adjacency-based problem of constructing contigs to the contig-based problem of constructing scaffolds. And as we shall see, since we are using statistics on gene adjacencies in evaluating reconstructions, an MWM approach feeds naturally into this step.

11.3 Excess Adjacencies as a Measure of Rearrangement

Independent rearrangements (inversions and reciprocal translocations) in newly diverging sister species with n genes tends to increase the total number of different adjacencies in the two genomes linearly at an initial rate of $2r$, where r is the total number of rearrangements in the two genomes ($r/2$ in each). At the same time, the number of adjacencies in common *decreases* at the same rate. If one of the genomes undergoes WGD soon after speciation or as part of speciation, the total number of different adjacencies in the two genomes still increases at a rate of $2r$. (The number of common adjacencies only decreases at a rate of r since rearrangement changes in the WGD descendant only affects one of the two identical adjacencies, leaving the other intact, so that only rearrangements in the other genome decreases the number of common adjacencies.) In either case—whether or not one of the genomes is a WGD descendant—the total number of different adjacencies in the two genomes, in excess of n, is an accurate measure of the degree of evolutionary divergence [8].

The advantage of this way of measuring evolutionary divergence over edit distances based on a repertoire of rearrangement operations, and over breakpoint distances, is that it applies equally well to comparing genomes with one or more WGD in their recent history as to those with no such history, and that it requires no special extension, constraint or modification wherever it is applied.

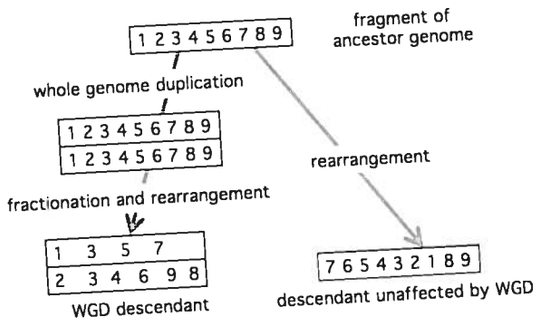

Fig. 11.1 Fractionation leading to different adjacencies in WGD descendant and unaffected genome. The adjacencies between genes 1 and 3, 3 and 5, 5 and 7 as well as 4 and 6 in the WGD descendant are caused by fractionation. The adjacency between 1 and 8 in the unaffected genome is caused by a reversal rearrangement, and the adjacency between genes 6 and 9 in the WGD descendant is caused by deletion of 7 and a rearrangement. Only two of the adjacencies are caused by rearrangement, but ignoring fractionation would lead to the inference of at least three more rearrangements to account for the different sets of adjacencies in the two genomes

11.4 Whole Genome Duplication and Fractionation

During fractionation, gene adjacency disruption follows from the random choice of which of the two copies is deleted, i.e., which copy of a chromosome retains the remaining single copy of the gene. This was first made explicit by Wolfe and Shields [17] in their original demonstration of "reciprocal gene loss" following the ancient WGD of *Saccharomyces cerevisiae*: "...this is the result of random deletion of individual duplicated genes from one or other chromosome subsequent to the initial duplication of the whole region." The pattern was further detailed later by the comparison of the *S. cerevisiae* gene order with that of related diploid yeasts [18, 19], where it was called"interleaving", while Freeling [5] coined the term "fractionation" in the context of plant genomics. Gordon et al. [20] and more recently, Ouangraoua et al. [21], have termed it "double synteny".

The phylogenomic extent of fractionation and the formal treatment of the deletion process have been the subject of numerous papers [22–26].

When a run of adjacent duplicate pairs lose a subset of their redundant genes from one chromosome and another, disjoint, subset from the other copy, as in Fig. 11.1, inference of the rearrangement distances between the WGD descendant and an unduplicated sister genome necessarily suggests that there are rearrangement breakpoints where adjacency no longer exists between the two subsets of single-copy survivors. This exaggerates the inferred number of reciprocal translocations and artificially inflates the overall amount of chromosomal rearrangement inferred between the two sister genomes.

We can correct this through the identification and isolation of "fractionation intervals", regions in both the WGD descendant and its unduplicated sister genome that have become partly or entirely single-copy in the former and may or may not

have been rearranged internally, but have (so far) been unaffected in both genomes by rearrangements exchanging genes from within the interval and genes external to the interval. The statistical properties of the intervals bear on current topics of interest in plant evolutionary genomics, whether duplicated genes are silenced or deleted one by one or through the deletion of longer stretches of DNA [22–24] and whether a fractionation regions tends to lose genes largely from one of the homeologous chromosomal segments or equally from the two [27].

11.5 Consolidation

Ideas about combining the information from the two fractionated regions in reconstructing ancestral genomes may be found in [5] for plants, in [20] for yeast and, more formally, in [21] for ancient vertebrates.

We have been developing a series of *consolidation algorithms* to identify and handle all instances of fractionation in a WGD descendant. The first of these [8] focuses on detecting and accounting for pairs of regions of single-copy genes in the WGD descendant that contain no genes in common (since the genes concerned are single-copy) but whose combined (or *consolidated* [5]) gene content is exactly the same as some contiguous region in a related genome unaffected by the WGD. A recent improvement in collaboration with Katharina Jahn and Jakub Kováč has linear run time, allows duplicate genes to be shared by the two intervals in the WGD descendant, and also extends the analysis to whole genome triplication and higher polyploidies [9]. Current work by Jahn solves the more difficult problem of comparing two fractionated sister genomes while dispensing with any necessity of referencing an unduplicated genome.

In the WGD case, once the pairs of regions or intervals are identified, together with the corresponding interval in the unduplicated genome, all three regions are replaced by a new, labeled, *virtual* gene.

The two genomes, thus altered by the creation of virtual genes replacing fractionated regions, are then examined for excess adjacencies and compared with the corresponding quantity in the untreated genome.

When one of the two intervals in the WGD descendant is empty because of completely biased fractionation, the corresponding virtual gene is still replaced in the appropriate context, deduced by examining the contexts of the other copy of the virtual gene in the WGD descendent and in the unaffected sister genome.

The consolidation algorithm treats the fractionation intervals as identical units, two in the WGD descendant and one in the unaffected genome. In this way it accounts for rearrangements which includes a whole interval in its scope, but also rearrangements which disrupt an interval, in that a fractionation involving such an interval will generally be automatically counted as two intervals, resulting in two virtual units instead of one. What the consolidation algorithm does not account for, however, are rearrangements occurring completely *within* one of the fractionation intervals.

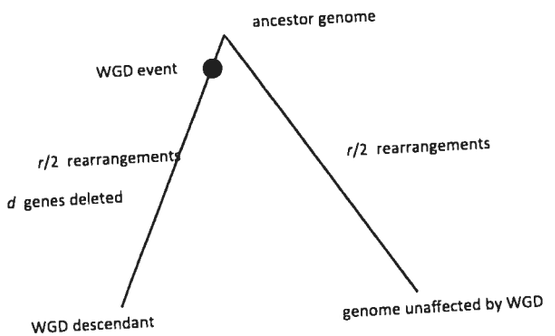

Fig. 11.2 Evolutionary scheme for simulating the production of excess adjacencies in a WGD descendant and an unaffected sister genome, by rearrangement and fractionation in the former, and rearrangement only in the latter. Ancestor contained 24,000 genes, divided among 20 chromosomes. Simulations carried out with number of random rearrangements (10 % reciprocal translocations, 90 % inversions) $r = 0, 600, 1800, 3000, 4200$ and random deletions $d = 0, 3000, \ldots, 21{,}000$

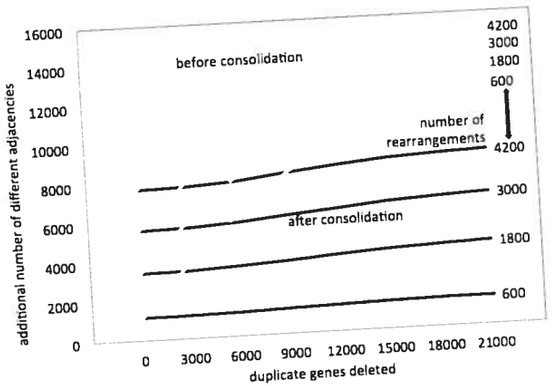

Fig. 11.3 Simulated effect of fractionation on increasing number of excess adjacencies before consolidation (*dashed lines*) and after (*solid lines*) for two genomes, one having undergone a WGD and one unaffected, diverging by various amounts of rearrangement

To correct for this, within each fractionation region, we first consider all the adjacencies in the three component intervals. We find an order for all the genes in the interval in the reconstructed ancestor genome, such that the following condition is satisfied. The induced sub-order determined by the subset of the genes from each of the three intervals in the extant genomes, two in the WGD descendant and one in the unaffected genome, has a minimum number of excess adjacencies when compared with the extant order, summed over all three intervals. We add the number of these interval-internal adjacencies to the set of adjacencies produced by the consolidation algorithm.

Figure 11.2 shows a scheme for simulating fractionation process following a WGD event. Figure 11.3, adapted from [8], shows how the consolidation algorithm wipes out almost all the bias caused by fractionation.

Table 11.1 Statistics on the reconstruction of the common ancestor of grape and poplar, before and after taking into account consolidation of the fractionation intervals. Of note is the decrease in the percentage of excess adjacencies (in boldface), representing artifactual rearrangements when fractionation is not taken into account

Genes in comparison	Grape	Poplar	Ancestor
Single copies	12,494	4,282	8,631
In syntenic pairs	0	$2 \times 8212 = 16,424$	0
Total	12,494	20,706	8,631
Adjacency statistics	**Before fractionation analysis**		
Adjacencies	12,475	20,676	8,588
Distinct (a)	12,475	16,165	8,588 (b)
Distinct overall (c)	19,446		
Excess (c–a)	6,971 (55.9 %)	3,281 (20.3 %)	
With ancestor (d)	9,390	11,094	**total**
Excess (d–b)	802 (9.3 %)	2,506 (**29.2 %**)	3,308 (**38.5 %**)
Virtual genes	**After fractionation analysis and consolidation**		
Fractionation intervals		2,462	1,888
Single copies	10,674	0	8143[a]
In syntenic pairs	0	$2 \times 10,674^{b} = 21,348$	0
Total	10,674	21,348	8143
Adjacency statistics	**After consolidation**		
Adjacencies	10,655	21,318	8,107
Distinct (a)	10,655	13,309	8,107 (b)
Distinct overall (c)	15,278		
Excess (c–a)	4,623 (43.4 %)	1,969 (14.8 %)	
With ancestor (d)	9,079	9,844	**total**
Excess (d–b)	972 (12.0 %)	1,737 (**21.4 %**)	2,709 (**33.4 %**)

[a]Counting genes within virtual genes: 9502

[b]Includes duplicate genes (not in a fractionation interval) and two copies of virtual genes even if only one gene-containing interval was found in poplar

11.6 Grape and Poplar

We applied our method to the genome of poplar (*Populus trichocarpa*) [28], which descends from a WGD event some 70 million years ago, and grape (*Vitis vinifera*) [29], which has undergone no WGD since the two genomes diverged some 130 million years ago.

As can be seen in Table 11.1, we discovered that a good proportion, over 25 %, of the apparent rearrangement in the poplar lineage, is actually attributable to fractionation. This is remarkable since only about 20 % of the poplar genome is made up of single-copy regions.

Another advantage of consolidation is that it resolves a major part of the "short contigs" problem of the MWM approach. The first stage of the MWM in the reconstruction before consolidation produced 2598 contigs with 12,494 genes. But once we applied the consolidation algorithm only 967 contigs were produced by the MWM, lengthening the average contig size by a factor of 2.6.

The benefits were less striking but still non-negligible in the second stage MWM, itself designed to overcome the problem of short contigs. Here, instead of 43 scaffolds in the reconstruction before consolidation, seven additional "joins" appeared, for a reduction to 36 scaffolds in the consolidated data.

11.7 Simulations

In our complex model of genome divergence through rearrangement, WGD and fractionation can only be validated by seeing how many aspects of the simulated output genomes match those of the real genome, with a minimum of model parameters. The number of genes in the two genomes, and the number of single-copy genes are fixed quantities, determined by the real genomes. Rearrangement can be carried out by a mixture of $(1-\theta)\rho$ short inversions, where the number of genes in the scope of the inversion is geometrically distributed with mean μ, plus $\theta\rho$ unbounded rearrangements whose endpoints are chosen randomly on chromosomes. In each deletion event the number of contiguous genes lost is geometrically distributed with mean λ. Finally, we introduce a parameter π for fractionation bias, the probability that a deletion takes place in a specified "subgenome", one of the two original copies of the duplicated ancestral genome created by the WGD event. There are thus five parameters that must be set for each simulation, ρ, θ, μ, λ, π plus the given structure of the ancestral genome, determined in our case, by the number of homologs in the poplar and grape genomes, and the number of single-copy genes in poplar. To find the appropriate values of the parameters to simulate the data, we can observe the total number of adjacencies between the output genomes, both before consolidation (R_1) and after consolidation (R_2). We can measure the average size L of the fractionation intervals (or, equivalently, the number of intervals N, since the product of the two quantities is fixed). And we can also indirectly observe the fractionation bias P, which is the deviation from an even split of the deleted genes of from the two copes of the fractionation interval in poplar or its simulation. More specifically we can measure $P(1)$, $P(2)$, ... in pairs of poplar fractionation intervals totaling $1, 2, \ldots$ genes, respectively. We term this "indirect" since we do not have access in the real poplar genome to the identity of genes in terms of their origin in one of the other "subgenomes" produced by the WGD event. We simply measure how many more genes there are in the larger fractionation interval compared to its counterpart, a value that is larger, on the average than the "true" bias.

We carried out a cyclical search, one parameter at a time, to find settings (Table 11.2) that gave the same average R_1, R_2 and N over 50 simulations as the values

Table 11.2 Best parameter values

Parameters	ρ	θ	μ	λ	π
Best values	1570	0.05	2.47	1.32	0.70

Table 11.3 Simulation statistics compared to real genomes. In each case the standard deviation over 50 samples was less than 1 % of the mean of the variable

	Real genomes		50 simulations	
	Before Consolidation	After Consolidation	Before Consolidation	After Consolidation
Total adjacencies (R)	19,538	15,363	19,563	15,298
Fractionation intervals (N)		2,462		2,458

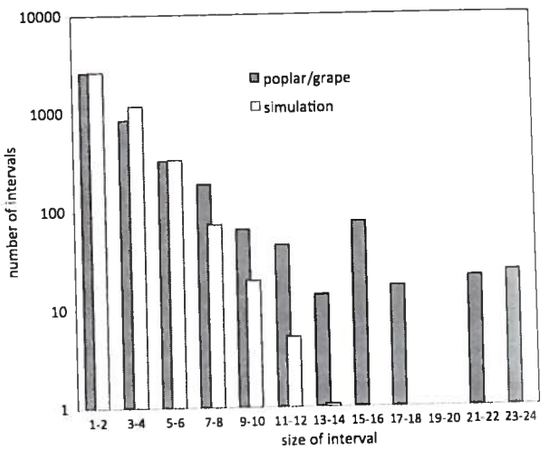

Fig. 11.4 Size of poplar/grape fractionation regions compared to simulations with parameters set so that reconstruction statistics match

calculated from grape and poplar (Table 11.3). We adjusted π so that the plot of the average simulated $P(i)$ resembled that from the real genomes.

Examining the consolidated regions detected by our algorithm, there are a number of regions much longer than those in the simulations (Fig. 11.4), suggesting a non-independence of deletion events affecting neighboring genes, and clear tendency for genes to be deleted in one of the two homeologs, as would be predicted by the recent theory of subgenome dominance [27].

In Fig. 11.5, a high value ($p = 0.85$) for the fractionation bias fits the data from poplar well only for long single-copy intervals, which are relatively rare (see Fig. 11.4), but a lower value ($p = 0.70$) fits the more numerous short intervals better. This is strong evidence that the selection of the various deletion sites does not proceed entirely independently, but rather that there are some regions of the genome that are particularly prone to becoming single-copy.

There are more parameters (five) to set in the simulations than quantities to observe (four) (though P is a vector, we can only observe its trend with any accuracy,

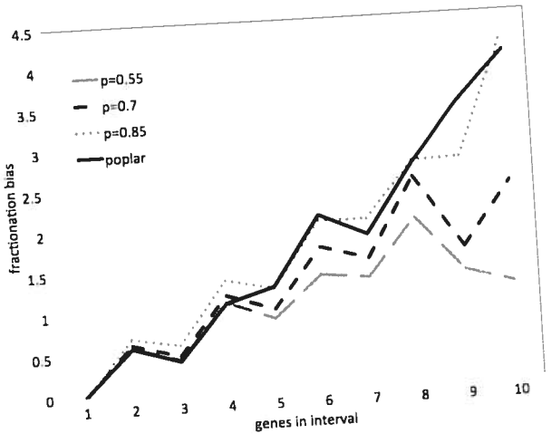

Fig. 11.5 Discrepancy in pairs of poplar intervals in the number of genes, compared with simulations with fractionation bias $p = 0.55, 0.70, 0.85$. Deletion event size geometrically distributed with mean 1.3. Jagged nature of graphs due not to statistical fluctuation but to measurement of discrepancy from an "even" split, which is necessarily calculated slightly differently for even and odd totals of genes in the fractionation intervals

not the individual $P(i)$), so there is inherently some non-uniqueness associated with the best choice of parameters, as suggested in Fig. 11.6. Nevertheless, the parameters can only take on values in a very restricted region. For example, outside a narrow range ρ produces too few or too many adjacencies, no matter what the settings are for the other parameters. And given ρ, the number of intervals N is sensitive to both parameters μ and λ.

11.8 Conclusions

The most important result from this work is that consolidation has the effect of greatly increasing the length of ancestral contigs output from the first MWM stage. The scaffolding approach already compensates for the short contigs problem, but combining the two strategies yields even longer scaffolds.

We have shown that we can closely simulate the gene order evolution of a WGD descendant and an unaffected sister genome, lending some confidence to our reconstruction of their common ancestor. We do detect, however, a significant number of long single-copy intervals, with highly biased fractionation, in the poplar genome, lying well outside the scope of our simulations. Whether there are biological connections among the genes in these intervals, and whether there are genes with no detected homologies in grape that are also present as single copies in these intervals, are questions for further study.

Further work will also involve improvements in parameter estimation as well as the identification of other measurable properties of evolutionary scenarios, restoring the balance between the number of parameters and the number quantities observed, in order to dispel problems of non-uniqueness.

This work has been undertaken as part of a project to formally analyze aspects of WGD fractionation, especially in the context of angiosperm evolution. Other directions include allowing duplicates in pairs of fractionation intervals, treating ploidies

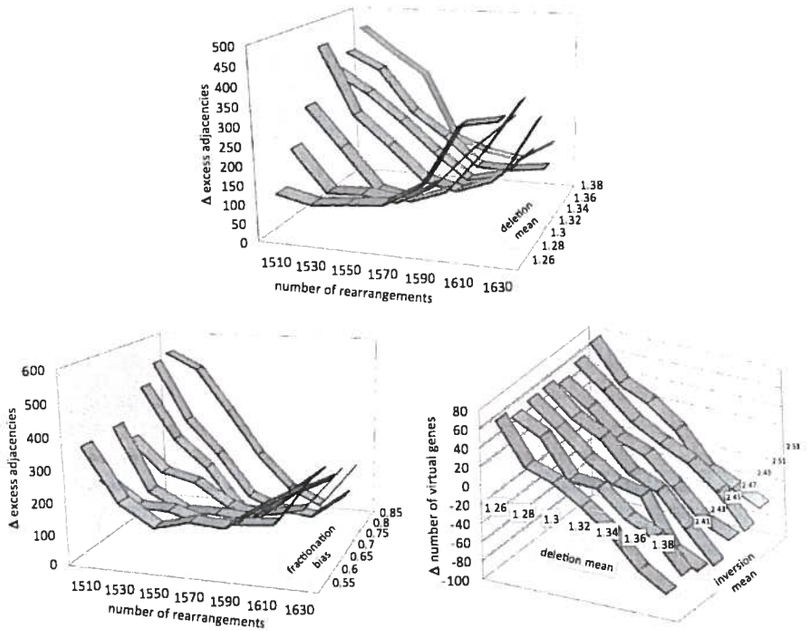

Fig. 11.6 Non-independent effects of simulation parameters on reconstruction characteristics. *Top*: Difference between real and simulated genomes (sum of R_1 and R_2 absolute differences) as a function of ρ and λ, showing the dependence of minimizing values of the parameters. *Left*: Difference between real and simulated genomes as a function of ρ and π, showing the dependence of minimizing values of the parameters. *Right*: Difference between real and simulated genomes (number of virtual genes) as a function of λ and μ, showing the *independence* of minimizing values of these particular parameters

of higher degree than WGD, dispensing with the necessity of an unaffected sister genome, as well as a probabilistic model of the distribution of fractionation interval sizes.

Acknowledgements We thank Katharina Jahn for many valuable comments and suggestions during the work reported here, and Vic Albert and Eric Lyons for guidance to current trends in angiosperm genomics. 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.

References

1. Chauve, C., Tannier, E.: A methodological framework for the reconstruction of contiguous regions of ancestral genomes and its application to mammalian genomes. PLoS Comput. Biol. **4**, 11 (2008)

2. Alekseyev, M.A., Pevzner, P.A.: Breakpoint graphs and ancestral genome reconstructions. Genome Res. **19**, 943–957 (2009)
3. Gagnon, Y., Blanchette, M., El-Mabrouk, M.: A flexible ancestral genome reconstruction method based on gapped adjacencies. BMC Bioinform. **13**(S19), S4 (2012)
4. Zheng, C., Chen, E., Albert, V.A., Lyons, E., Sankoff, D.: Ancient eudicot hexaploidy meets ancestral eurosid gene order. BMC Genomics **14** (2013, in press)
5. Langham, R.J., Walsh, J., Dunn, M., Ko, C., Goff, S.A., Freeling, M.: Genomic duplication, fractionation and the origin of regulatory novelty. Genetics **166**, 935–945 (2004)
6. Soltis, D.E., Albert, V.A., Leebens-Mack, J., Bell, C.D., Paterson, A.H., Zheng, C., Sankoff, D., dePamphilis, CW, Wall, P.K., Soltis, P.S.: Polyploidy and angiosperm diversification. Am. J. Bot. **96**, 336–348 (2009)
7. El-Mabrouk, N.: Genome rearrangement by reversals and insertions/deletions of contiguous segments. In: Giancarlo, R., Sankoff, D. (eds.) Combinatorial Pattern Matching (CPM 2000), Proceedings of the 11th Annual Symposium. Lecture Notes in Computer Science, vol. 1848, pp. 222–234 (2000)
8. Sankoff, D., Zheng, C.: Fractionation, rearrangement and subgenome dominance. Bioinformatics **28**, 402–408 (2012)
9. Jahn, K., Zheng, C., Kováč, J., Sankoff, D.: A consolidation algorithm for genomes fractionated after higher order polyploidization. BMC Bioinform. **13**(S19), S8 (2012)
10. Lyons, E., Freeling, M.: How to usefully compare homologous plant genes and chromosomes as DNA sequences. Plant J. **53**, 661–673 (2008). http://genomevolution.org/CoGe/
11. Lyons, E., Pedersen, B., Kane, J., Alam, M., Ming, R., Tang, H., Wang, X., Bowers, J., Paterson, A., Lisch, D., Freeling, M.: Finding and comparing syntenic regions among Arabidopsis and the outgroups papaya, poplar and grape: CoGe with rosids. Plant Physiol. **148**, 1772–1781 (2008)
12. Zheng, C., Swenson, K., Lyons, E., Sankoff, D.: OMG! Orthologs in multiple genomes—competing graph-theoretical formulations. In: Przytycka, T.M., Sagot, M.-F. (eds.) Algorithms in Bioinformatics, Proceedings of the 11th International Workshop. Lecture Notes in Computer Science, vol. 6833, pp. 364–375 (2011)
13. Galil, Z.: Efficient algorithms for finding maximum matching in graphs. ACM Comput. Surv. **18**, 23–38 (1986)
14. Tannier, E., Zheng, C., Sankoff, D.: Multichromosomal median and halving problems under different genomic distances. BMC Bioinform. **10**, 120 (2009)
15. Warren, R., Sankoff, D.: Genome aliquoting revisited. J. Comput. Biol. **18**, 1065–1075 (2011)
16. Manuch, J., Patterson, M., Wittler, R., Chauve, C., Tannier, E.: Linearization of ancestral multichromosomal genomes. BMC Bioinform. **13**(S19), S11 (2012)
17. Wolfe, K.H., Shields, D.C.: Molecular evidence for an ancient duplication of the entire yeast genome. Nature **387**, 708–713 (1997)
18. Dietrich, F.S., Voegeli, S., Brachat, S., Lerch, A., Gates, K., Steiner, S., Mohr, C., Pöhlmann, R., Luedi, P., Choi, S., Wing, R.A., Flavier, A., Gaffney, T.D., Philippsen, P.: The Ashbya gossypii genome as a tool for mapping the ancient Saccharomyces cerevisiae genome. Science **304**, 304–307 (2004)
19. Kellis, M., Birren, B.W., Lander, E.S.: Proof and evolutionary analysis of ancient genome duplication in the yeast Saccharomyces cerevisiae. Nature **428**, 617–624 (2004)
20. Gordon, J.L., Byrne, K.P., Wolfe, K.H.: Additions, losses, and rearrangements on the evolutionary route from a reconstructed ancestor to the modern Saccharomyces cerevisiae genome. PLoS Genet. **5**, e1000485 (2009)
21. Ouangraoua, A., Tannier, E., Chauve, C.: Reconstructing the architecture of the ancestral amniote genome. Bioinformatics **27**, 2664–2671 (2011)
22. Byrnes, J.K., Morris, G.P., Li, W.-H.: Reorganization of adjacent gene relationships in yeast genomes by whole-genome duplication and gene deletion. Mol. Biol. Evol. **23**, 1136–1143 (2006)
23. van Hoek, M.J., Hogeweg, P.: The role of mutational dynamics in genome shrinkage. Mol. Biol. Evol. **24**, 2485–2494 (2007)

24. Sankoff, D., Zheng, C., Zhu, Q.: The collapse of gene complement following whole genome duplication. BMC Genomics **11**, 313 (2010)
25. Wang, B., Zheng, C., Sankoff, D.: Fractionation statistics. BMC Bioinform. **12**(S9), S5 (2011)
26. Sankoff, D., Zheng, C., Wang, B.: A model for biased fractionation after whole genome duplication. BMC Genomics **13**(S1), S8 (2012)
27. Schnable, J., Springer, N., Freeling, M.: Differentiation of the maize subgenomes by genome dominance and both ancient and ongoing gene loss. Proc. Natl. Acad. Sci. USA **108**, 4069–4074 (2011)
28. Tuskan, G.A., Difazio, S., Jansson, S., Bohlmann, J., Grigoriev, I., Hellsten, U., Putnam, N., Ralph, S., Rombauts, S., Salamov, A. et al.: The genome of Black Cottonwood, Populus trichocarpa (Torr. & Gray). Science **313**, 1596–1604 (2006)
29. Jaillon, O., Aury, J.M., Noel, B., Policriti, A., Clepet, C., Casagrande, A., Choisne, N., Aubourg, S., Vitulo, N., Jubin, C., et al.: The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. Nature **449**, 463–467 (2007)

Chapter 12
Error Detection and Correction of Gene Trees

Manuel Lafond, Krister M. Swenson, and Nadia El-Mabrouk

Abstract Reconstructing the phylogeny of a gene family and reconciling the obtained gene tree with the species tree reveals the history of duplications, losses, and other events that have shaped the gene family, with important implications towards the functional specificity of genes. However, evolutionary histories inferred by reconciliation are strongly dependent upon the accuracy of the trees, and few misplaced leaves will lead to a completely different history. Furthermore, sequence data alone often lack the information to confidently support a gene tree topology. We outline a number of criteria that can be used to detect erroneous gene trees. Analysing *Ensembl* gene trees of the fish genomes Stickleback, Medaka, Tetraodon, and Zebrafish reveals a significant number of erroneous gene trees. Finally, some potential directions for error correction of gene trees are explored.

12.1 Introduction

Duplication followed by modification is a major mechanism driving evolution. Consequently, genes cannot be seen as independent entities, but rather as entities related through duplication and speciation events. Grouping genes into families of *homologs* (i.e. copies originating from a single ancestral gene) and reconstructing the phylogeny of each gene family is requisite for a variety of annotation, evolutionary, and functional studies. By *reconciling* such a gene tree with a species tree, one can infer the history of duplications, losses and other events that have shaped the gene family. Such a history reveals the orthology (evolution of the ancestral

M. Lafond (✉) · K.M. Swenson · N. El-Mabrouk
Département d'Informatique et de Recherche Opérationnelle (DIRO), Université de Montréal, CP 6128, Succursale Centre-ville, Montréal, QC H3C 3J7, Canada
e-mail: lafonman@iro.umontreal.ca

N. El-Mabrouk
e-mail: mabrouk@iro.umontreal.ca

K.M. Swenson
McGill University, Montreal, QC, Canada
e-mail: swensonk@iro.umontreal.ca

C. Chauve et al. (eds.), *Models and Algorithms for Genome Evolution*, Computational Biology 19, DOI 10.1007/978-1-4471-5298-9_12,