Internal Validation of Ancestral Gene Order Reconstruction in Angiosperm Phylogeny

David Sankoff[1], Chunfang Zheng[1], P. Kerr Wall[2], Claude dePamphilis[2], Jim Leebens-Mack[3], and Victor A. Albert[4]

[1] Dept. of Mathematics & Statistics and Dept. of Biology, University of Ottawa, Ottawa, ON, Canada K1N 6N5
{sankoff,czhen033}@uottawa.ca
[2] Biology Department, Penn State University, University Park, PA 16802, USA
{pkerrwall,cwd3}@psu.edu
[3] Department of Plant Biology, University of Georgia, Athens, GA 30602, USA
jleebensmack@plantbio.uga.edu
[4] Department of Biological Sciences, SUNY Buffalo, Buffalo, NY 14260, USA
vaalbert@buffalo.edu

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.

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 has undergone whole genome duplication [3,11,14] 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*, that have been sequenced to date, shown in Figure 1.

C.E. Nelson and S. Vialette (Eds.): RECOMB-CG 2008, LNBI 5267, pp. 252–264, 2008.

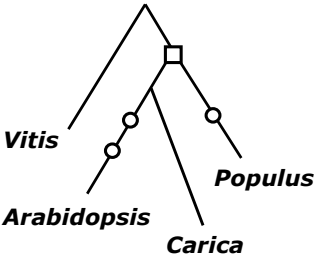

Fig. 1. Phylogenetic relationships among angiosperms with sequenced genomes. The circles indicate likely whole genome doubling events. The circle in the *Populus* lineage, representing the locus of the WGD event at the origin of the willow-poplar family, and the square, representing the ancestor of the 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 [4]. 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 [18,10,17,19,20], 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., those 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 that 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 [4]. 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 [17,20]. 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 [12]. In the heuristics necessary for these two approaches, there is a trade-off between the speed and quality of constrained halving versus the unbiased and possibly 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 rearranged genomes are algorithmically reconstructed, however, some breakpoints may be reused. If d is the number of rearrangements and b the number of breakpoints, the reuse [6] 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 [9]. This is particularly likely to occur for genomes that are only very distantly related.

Dispersion. The motivation for guided halving is to resolve the ambiguities inherent in the large number of solutions. One way to quantify the remaining non-uniqueness is to calculate the distances among a sample of solutions.

In this paper we will refer repeatedly to a main tabulation of results, Table 1, in which we discover the unexpected rapid evolution of the *Carica* gene order in comparison with that of *Vitis*. In Section 2, we report on the origin and processing of our gene-order data and the construction of the full and defective homology sets. Then, in Section 3, we discuss the formulation of genomic distances and the halving problems, and sketch a new algorithm for unconstrained guided halving. In Section 4 we evaluate the utility of singletons and of defective homology sets. Then, in Section 5 we assess the two guided halving algorithms on real and simulated data. Section 6 proposes a way to use unassembled genome sequence in contig form as input to the reconstruction algorithm, an approach that could potentially have wide use in gene order phylogeny. In Section 7 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, this was dictated by a paucity of data in the appropriate configurations.

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 [14], the French National Sequencing Center, Genoscope [5], and the University of Hawaii [8], respectively. An all-by-all BLASTP search was run on a data set including all *Populus* and *Vitis* protein coding genes, and orthoMCL

[7] 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.

3 Genome Distance, Breakpoint Graph, Guided Halving

Genome comparison algorithms generally involve manipulations of the bicoloured breakpoint graph [1,13] 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 [16].

Then the genome halving problem [4] 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 $A \oplus A$.

The guided genome halving problem [18] 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 [18,10,17], 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 halving problem and $d(T, A \oplus A) + d(A, R)$ is minimal. This has the advantage that a good proportion of the computation, namely the halving aspect, is guaranteed to be rapid and exact, although the overall algorithm, which is essentially a search among all optimal A, remains heuristic. Without sketching out the details of the lengthy algorithm, the addition of gray edges representing genome A to the breakpoint graph, as in Figure 2, must favour

Table 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.

	genes in A,	$d(A, Vitis)$			$d(A \oplus A, Populus)$			total
data sets	with singletons	d	b	r	d	b	r	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
	genes in A,	$d(A, Carica)$			$d(A \oplus A, Populus)$			total
data sets	with singletons	d	b	r	d	b	r	d
Solutions constrained to also be solutions of genome halving								
PPC	2590	896	1152	1.56	565	823	1.37	1461
PPC, PP	3478	905	1158	1.56	884	1282	1.38	1789
PPC,PC	6334	1892	2314	1.64	1262	1700	1.48	3154
PPC,PP,PC	7222	1925	2341	1.64	1541	2065	1.49	3466
Solutions unconstrained								
PPC	2590	864	1125	1.54	628	870	1.44	1492
PPC, PP	3478	873	1125	1.55	951	1318	1.44	1824
PPC,PC	6334	1859	2277	1.63	1321	1742	1.52	3180
PPC,PP,PC	7222	1877	2313	1.62	1617	2126	1.52	3494
	genes in A,	$d(A, Vitis)$			$d(A \oplus A, Populus)$			total
data sets	without singletons	d	b	r	d	b	r	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
	genes in A,	$d(A, Carica)$			$d(A \oplus A, Populus)$			total
data sets	without singletons	d	b	r	d	b	r	d
Solutions constrained to also be solutions of genome halving								
PPC	2464	772	1014	1.52	412	607	1.36	1184
PPC, PP	3226	812	1058	1.53	536	809	1.33	1348
PPC,PC	4651	779	1054	1.48	774	1050	1.47	1554
PPC,PP,PC	5234	898	1206	1.49	892	1253	1.42	1790
Solutions unconstrained								
PPC	2464	758	1001	1.51	454	639	1.42	1212
PPC, PP	3226	796	1046	1.52	584	839	1.39	1380
PPC,PC	4651	764	1041	1.47	804	1090	1.48	1568
PPC,PP,PC	5234	861	1178	1.46	952	1303	1.46	1813

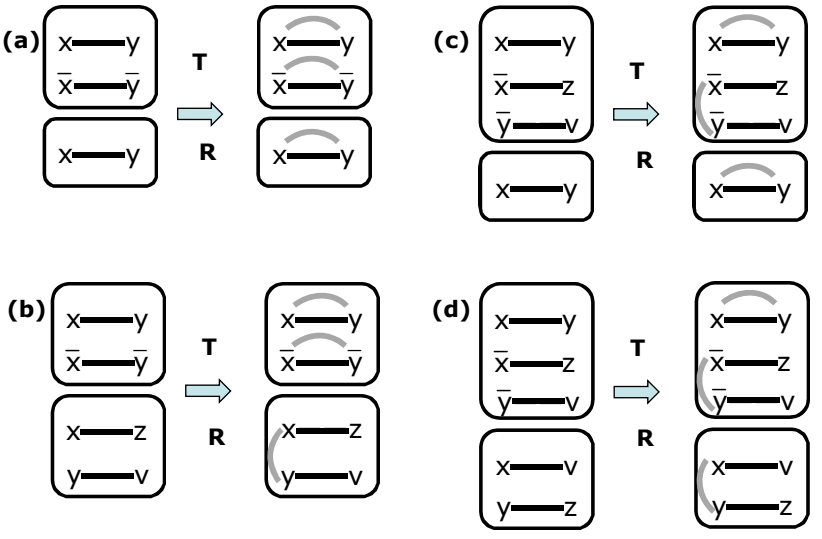

Fig. 2. Choice of gray edge to add at each stage of the reconstruction of A and $A \oplus A$. Each black edge in the diagram represents either an adjacency in T or R or an alternating colour path with a black edge at each end point. If vertex w is copy "1" in T then $\bar{w}$ is copy "2", and vice versa. (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).

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. [4]. 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 3.1, 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.

3.1 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 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 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 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 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. [4] 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.

4 On the Utility of Singletons and Defective Homology Sets

From the last column of Table 1, it is clear that of the four factors, inclusion/exclusion of singletons, inclusion/exclusion of defective homology sets, outgroup species and heuristic, the largest effects on total genomic distance are due

to the choice of homology sets and inclusion of singletons, while the heuristic used has a much smaller effect. We will return to the differences between the algorithms in Section 5, and to the choice of outgroup in Section 7, but we can observe here that the inclusion of the homology sets defective by virtue of one missing *Populus* copy increases the genomic distances disproportionately and also reduces the quality of the inference, as measured by r in all the analyses containing singletons, and all the *Populus*-$A \oplus A$ comparisons.

At the same time the inclusion of singletons had a major effect on the distance, especially where the PV or PC homology sets are included. In addition, by comparing all the sub-tables with singletons, in the top half of the table, with the corresponding sub-table without singletons, in the bottom half, the inclusion of singletons degrades the analysis, with few exceptions, as measured by an increase in the two r statistics, the one pertaining to the duplicated genome and the one pertaining to the outgroup.

5 Comparison of the Heuristics

In Table 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 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.

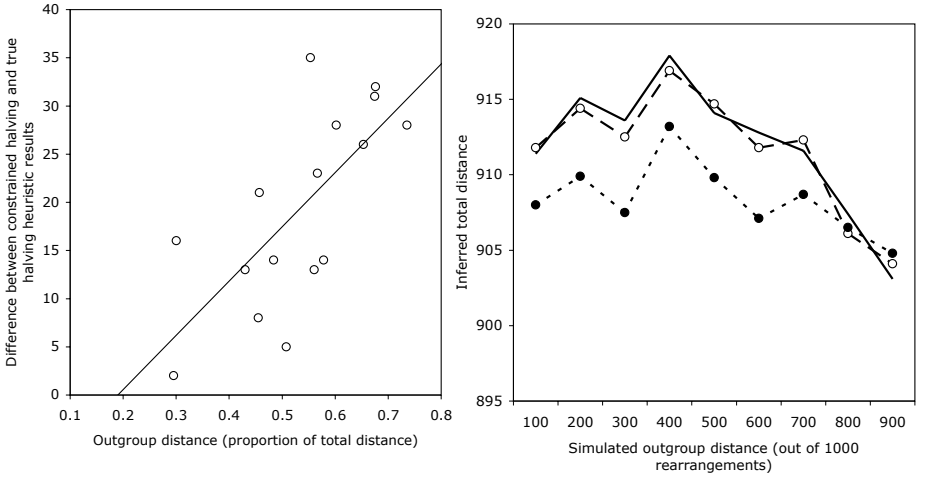

Fig. 3. Performance of the constrained and unconstrained heuristics as a function of the real (left) or simulated (right) distance of the outgroup from A

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 a genome R and $1000 - M$ rearrangements from genome $A \oplus A$ to produce a genome T. We then applied the constrained and the new algorithms, showing that the new one was superior when $M < 800$, but not for $M \geq 1000$, as seen in Figure 3 (right). Considering the 16 comparisons 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 3 (left).

The difference between the simulations, where the new method is always superior, and the real analysis, where the new method would seem to be superior only when the outgroup is very close to the ancestor, must be ascribed to some way the model used for the simulations does not fit the data. One clue is the relatively high reuse rate in the comparison between the outgroup and A, compared with that between *Populus* and $A \oplus A$.

6 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 [16], we may 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 2 illustrate how the $d(A, Carica)$ results in the bottom quarter of Table 1 were obtained.

Table 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, $d - c + 9$: distance corrected for excess of contigs over true number of chromosomes, a: number of 'obvious fusions". Data without singletons. Solutions obtained by constrained algorithm.

		$d(A, Carica)$			correction				
data sets	genes in A	d	b	uncorrected r	c	$d - c + 9$	a	$b - a$	corrected r
PPC	2464	986	1090	1.81	223	772	76	1014	1.52
PPC, PP	3226	1027	1132	1.81	224	812	74	1058	1.53
PPC,PC	4651	1084	1177	1.84	314	779	123	1054	1.48
PPC,PP,PC	5234	1214	1318	1.84	325	898	112	1206	1.49

Figure 4 (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 %.

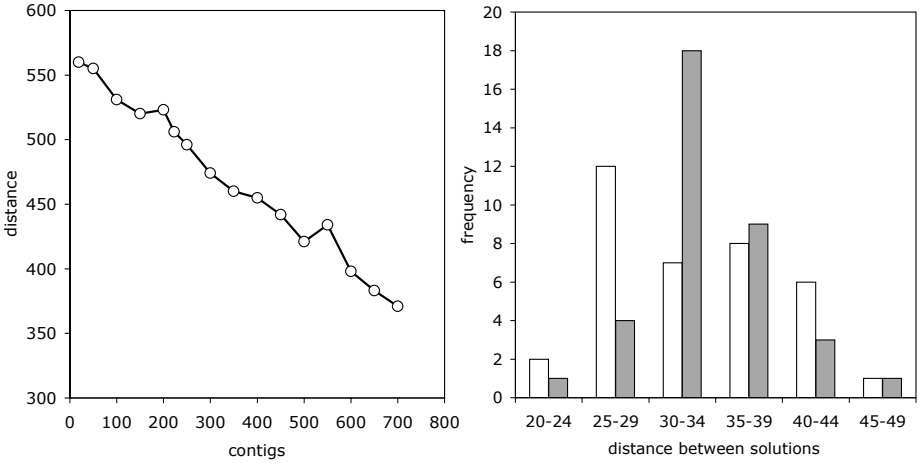

Fig. 4. Left: Effect of increasing fragmentation of *Vitis* into "contigs" on the distance between the reconstructed *A* and *Vitis*. Right: Distributions of distances among solutions for *A* based on *Vitis* data (white bars) and among solutions for *Vitis* fragmented into contigs in different random ways (gray bars).

Figure 4 (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 alternate solutions for the original *Vitis* data, a few percentage points of the distance itself.

7 A Comparison of the Outgroups

Perhaps the most surprising result of this study 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 Vitaceae family from the rosids, though some older taxonomies do not make this distinction.

Before interpreting this result, we should correct 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. This overestimation of the *Carica*-ancestor distance might account for about half the difference in the distances. But the other source of error is due to the contig data, and this results in an *underestimate* of the *Carica*-ancestor distance. From Figure 4, we can estimate that the *Carica* distances are underestimated by about 10 % because of the 223 contigs in the *Carica* data. So this increases the discrepancy between the two outgroups, restoring it almost to what it was before the corrections.

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*.

7.1 Using Both Outgroups

There are 1734 complete homologous gene sets including two *Populus* copies and one copy in each of *Carica* and *Vitis*. In the same way that the unconstrained algorithm in Section 3 is based on a modification of the guided halving algorithm for one outgroup in reference [17], 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 [17] to find the ancestor (small circle) genome in Figure 5(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 [12] and solving it is barely feasible with the 1734 genes in our data, requiring some 300 hours of MacBook computing time.

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". And in the same way, we may correct it by subtracting the number of contigs in excess of a reasonable number of chromosomes (19 in the median) from the distance in order to obtain a corrected distance. This corresponds to disregarding the fusions counted in the original distance that are essentially carrying out an optimal assembly, modeling an analytical process, not a biological one. This produces the corrected values in Figure 5(b).

Let us compare the distance from *Vitis* and from *Carica* to ancestor A, passing through the median, in Figure 5 (517 and 577, respectively), with the minimum distances[1] in Table 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 36 and by 34, 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 $481+170+543+170 = 1364$ to $517+577+170 = 1264$.

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

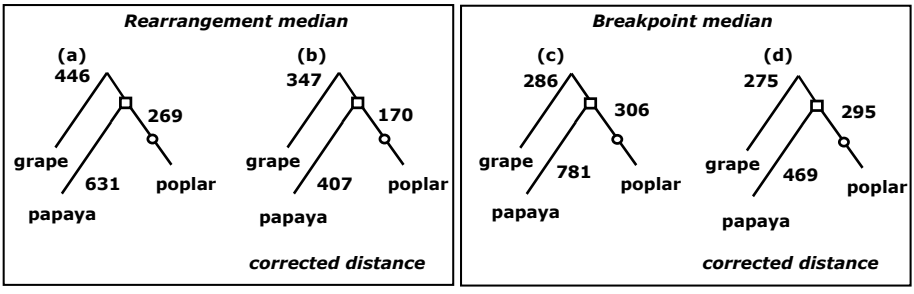

Fig. 5. Branch lengths in angiosperm phylogeny, using two estimates of the median, and applying the contig correction

There is one version of guided halving that is of polynomial complexity [12]. This involves a "general breakpoint model" for multichromosomal genomes, which does not explicitly refer to rearrangements. Running this algorithm, requiring only 15 MacBook minutes, on the three angiosperm genomes results in a median with only 30 contig-chromosomes. Calculating the rearrangement distances from this median to ancestor *A*, *Carica* and *Vitis* gives the results in Figure 5(c); correcting them for excess contigs gives the results in Figure 5(d).

Figures 5(b) confirm that the papaya genome has evolved more rapidly than the grapevine one. Figure 5(d) shows an even greater distance, although this is not based on the rearrangement median.

8 Conclusions

The main contributions of this paper are:

- The discovery of the rapid rate of gene order evolution in *Carica* compared to *Vitis*,
- 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 defective homology sets and 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.

References

1. Bafna, V., Pevzner, P.: Genome rearrangements and sorting by reversals. SIAM Journal of Computing 25, 272–289 (1996)
2. Bergeron, A., Mixtacki, J., Stoye, J.: A unifying view of genome rearrangements. In: Bücher, P., Moret, B.M.E. (eds.) WABI 2006. LNCS (LNBI), vol. 4175, pp. 163–173. Springer, Heidelberg (2006)

3. Cui, L., Wall, P.K., Leebens-Mack, J.H., Lindsay, B.G., Soltis, D.E., Doyle, J.J., Soltis, P.S., Carlson, J.E., Arumuganathan, K., Barakat, A., Albert, V.A., Ma, H., de Pamphilis, C.W.: Widespread genome duplications throughout the history of flowering plants. Genome Research 16, 738–749 (2006)
4. El-Mabrouk, N., Sankoff, D.: The reconstruction of doubled genomes. SIAM Journal on Computing 32, 754–792 (2003)
5. Jaillon, O., Aury, J.M., Noel, B., Policriti, A., Clepet, C., et al.: The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. Nature 449, 463–467 (2007), http://www.genoscope.cns.fr/externe/English/Pro-jets/Projet_ML/data/annotation/
6. Pevzner, P.A., Tesler, G.: Human and mouse genomic sequences reveal extensive breakpoint reuse in mammalian evolution. Proceedings of the National Academy of Sciences USA 100, 7672–7677 (2003)
7. Li, L., Stoeckert Jr., C.J., Roos, D.S.: OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Research 13, 2178–2189 (2003)
8. Ming, R., Hou, S., Feng, Y., Yu, Q., Dionne-Laporte, A., et al.: The draft genome of the transgenic tropical fruit tree papaya (Carica papaya Linnaeus). Nature 452, 991–996 (2008), http://asgpb.mhpcc.hawaii.edu
9. Sankoff, D.: The signal in the genomes. PLoS Computational Biology 2, e35 (2006)
10. Sankoff, D., Zheng, C., Zhu, Q.: Polyploids, genome halving and phylogeny. Bioinformatics 23, i433–i439 (2007)
11. Soltis, D., Albert, V.A., Leebens-Mack, J., Bell, C.D., Paterson, A., Zheng, C., Sankoff, D., dePamphilis, C.W., Wall, P.K., Soltis, P.S.: Polyploidy and angiosperm diversification. American Journal of Botany (in press, 2008)
12. Tannier, E., Zheng, C., Sankoff, D.: Multichromosomal median and halving problems under different genomic distances. In: Workshop on Algorithms in Bioinformatics WABI 2008 (in press, 2008)
13. Tesler, G.: Efficient algorithms for multichromosomal genome rearrangements. Journal of Computer and System Sciences 65, 587–609 (2002)
14. Tuskan, G.A., Difazio, S., Jansson, S., Bohlmann, J., Grigoriev, I., et al.: The genome of black cottonwood, Populus trichocarpa (Torr. & Gray). Science 313, 1596–1604 (2006)
15. Velasco, R., Zharkikh, A., Troggio, M., Cartwright, D.A., Cestaro, A., et al.: A high quality draft consensus sequence of the genome of a heterozygous grapevine variety. PLoS ONE 2, e13–e26 (2007)
16. Yancopoulos, S., Attie, O., Friedberg, R.: Efficient sorting of genomic permutations by translocation, inversion and block interchange. Bioinformatics 21, 3340–3346 (2005)
17. Zheng, C., Zhu, Q., Adam, Z., Sankoff, D.: Guided genome halving: hardness, heuristics and the history of the Hemiascomycetes. Bioinformatics 24, i96–i104 (2008)
18. Zheng, C., Zhu, Q., Sankoff, D.: Genome halving with an outgroup. Evolutionary Bioinformatics 2, 319–326 (2006)
19. Zheng, C., Zhu, Q., Sankoff, D.: Descendants of whole genome duplication within gene order phylogeny. Journal of Computational Biology 15 (in press, 2008)
20. Zheng, C., Wall, P.K., Leebens-Mack, J., Albert, V.A., dePamphilis, C.W., Sankoff, D.: The effect of massive gene loss following whole genome duplication on the algorithmic reconstruction of the ancestral Populus diploid. In: Proceedings of the International Conference on Computational Systems Bioinformatics CSB 2008 (in press, 2008)