

Inferring Ancestral Chloroplast Genomes with Inverted Repeat

Liyong Cui[‡], Feng Yue[†], Claude W. dePamphilis[‡], Bernard M.E. Moret^{*}, Jijun Tang[†]

[‡]Department of Biology, Pennsylvania State University,
University Park, PA 16802, USA

[†]Dept. of Computer Science & Engineering, University of South Carolina,
Columbia, SC 29208

^{*}Department of Computer Science, University of New Mexico,
Albuquerque, NM 87131

Abstract— Genome evolution is shaped not only by nucleotide substitutions, but also by structural changes including gene and genome duplications, insertions/deletions and gene order rearrangements. Reconstruction of phylogeny based on gene order changes has been limited to cases where equal gene content or few deletions can be assumed. Since conserved duplicated regions are present in many Chloroplast genomes, the inference of inverted repeats is needed in ancestral genome reconstruction. We apply GRAPPA-IR to reconstruct ancestral chloroplast genomes containing duplicated genes. A test of GRAPPA-IR using divergent chloroplast genomes from land plants and green algae recovers the phylogeny congruent with prior studies, while analysis that do not consider IR structure fail to obtain the accepted topology. The ancestral genome structure suggests that genome rearrangement in chloroplasts is probably limited by inverted repeats with a conserved core region. In addition, the boundaries of inverted repeats are hot spots for gene duplications or deletions.

Keywords: Phylogeny, Genome Rearrangement, Chloroplast, Inverted Repeat, Inversion.

1. Introduction

Mutations in a genome consist of not only base pair level changes but also events that alter the chromosome structure, such as inversions, duplications or deletions [13]. Ancestral gene sequence inference has led to significant predictions of protein functional shift and positive selection [19]. For example, comparisons of orthologous chromosomal segments showed heterogeneous rates of evolution of the X chromosome in human, mouse and rat [8]. However, on the genome level, the evolutionary change of genome structure is poorly understood. Inference of ancestral genomes was mainly achieved at the DNA level, but limited to closely related

organisms where rearrangements were negligible, partly because of the complexity in assigning genes in duplicated segments to orthologous groups [2].

In this paper, we take a simple, alternative data set of chloroplast genomes to study the genome structural changes. Chloroplasts are the green, photosynthetic organelles that originated from a free-living cyanobacteria-like ancestor [27]. Chloroplasts maintained a reduced genome through over one billion years of endosymbiosis [15]. Typical chloroplast genomes are circular single chromosomes with 120 – 200 genes, which encode proteins, tRNAs, rRNAs and hypothetical open reading frames. Most chloroplast genomes consist of four distinct parts: two duplicated regions (inverted repeats, or *IR*) separated by a large single copy (*LSC*) and a small single copy (*SSC*) region. One common characteristic of the chloroplast *IR* is the presence of three rRNA genes (*rnn5s*, *rnn16s* and *rnn23s*, or *rrf*, *rrs*, and *rri*), which are homologous to genes of the cyanobacteria *rnn* operon. The structure of chloroplast genomes of land plants is highly conserved, with almost collinear gene order, except for elevated level of rearrangements in specific lineages including green algae [16], conifers and members of the flowering plant families *Campanulaceae* [5], *Geraniaceae* [26] and *Fabaceae* [23]. The gene content of chloroplast *IRs* vary greatly, largely due to the expansion and contraction of the *IR* at the *IR-SC* boundaries; this “ebb and flow” of *IR* boundary has been observed even within a genus [9], [25]. Chloroplast genomes of green algae (charophyte and chlorophyte algae) also contain more variations of gene order and some are highly rearranged [16]. Because of their compact size and the availability of conserved DNA probes,

many chloroplast genomes have been mapped [6], and 42 have been completely sequenced to date. Thus, chloroplast genomes provide an ideal example for modeling genome rearrangements over a broad evolutionary time scale.

Genomes are thought to evolve by inversions and transpositions, as well as deletions, insertions and duplications of fragments. For example, Let G be the genome with signed ordering of (g_1, g_2, \dots, g_n) , an *inversion* between indices i and j ($i \leq j$) produces the genome with linear ordering as $(g_1, g_2, \dots, g_{i-1}, -g_j, -g_{j-1}, \dots, -g_i, g_{j+1}, \dots, g_n)$. Because gene rearrangements are *rare genomic events* [29], gene-order data enable the reconstruction of evolutionary events far back in time, thus many biologists have embraced this new source of data in their phylogenetic work [5], [6], [21].

Gene order phylogeny was first proposed by Sankoff and an algorithm using break-point distance was implemented in `BPAAnalysis` [2]. The inversion phylogeny was introduced to improve the accuracy and was implemented in `GRAPPA`. Extensive simulations showed that inversion median were superior to breakpoint median [18] and the trees returned were more accurate than using either distance-based or parsimony methods [17]. The current version of `GRAPPA` (version 2.0) is able to estimate the phylogeny and inversion medians using genomes with equal gene content (i.e., no insertion, deletion or duplication are allowed) [17]. A scaled-up version, `DCM-GRAPPA`, is able to estimate the gene-order phylogeny with very high accuracy for thousands of genomes, thus greatly increasing the power of genome phylogeny using large datasets [31].

Ancestral genome reconstruction was not formally validated using `GRAPPA` on true datasets. Since the inversion model is closer to the biological process of genome rearrangements, we can use the inferred internal genomes to estimate ancestral gene orders. Two challenges exist in designing such algorithm for ancestral chloroplast genome reconstruction: first, the algorithm needs to be accurate over heterogeneous branch lengths on a phylogeny since the rate of chloroplast genome rearrangements varies among lineages. Second, it should be able to deal with large segment of gene duplications and deletions that lead to the inverted repeats, while different genomes may have different set of genes in the IR region.

We extended `GRAPPA` [32] so that it is able to analyze data sets with a limited number of dele-

tions, but no duplication is allowed. To remedy this problem, we develop a new algorithm (`GRAPPA-IR`) for chloroplast genomes that take into account the unique quadripartite structure (e.g., LSC-IRa-SSC-IR), which is common to not only the chloroplast genomes, but also some other IR-containing DNAs. The assumption of our new approach is that inversions do not occur across inverted repeats, because the genome structure will be disrupted by such inversions that “flip” the repeats from inverted to the same orientation. According to the model, a change of gene content within the IR region is mainly due to growth or shrinkage of the IR at the IR-SC boundaries. This approach is in agreement with the observation that in most IR-containing chloroplasts, the gene content in the whole genome is conserved, but IR length and IR gene contents can be varied.

2. The Dataset

To better present the new `GRAPPA-IR` method, we select a test case of six chloroplast genomes representing major lineages of land plants and green algae, all of which share the quadripartite structure (LSC-IRa-SSC-IRb). The organisms include *Nicotiana tobacum* (tobacco, *nt*), *Psilotum nudum* (whisk fern, *pn*), *Marchantia polymorpha* (liverwort, *mp*), *Chaetosphaeridium globosum* (a charophyte alga, *cp*), *Nephroselmis olivacea* (a chlorophyte alga, *no*), and *Mesostigma viride* (a photosynthetic protist, *mv*). A reference phylogenetic tree was constructed using the maximum parsimony method with 50 concatenated proteins. Cyanophora proteins were used to root the tree (Fig. 1). The reference tree is the same as the phylogeny by Lemieux et al. [14] in which *Mesostigma* basal to other green plants. An alternative phylogeny was published by Karol et al. based on maximum likelihood analysis of four chloroplast genes and including many more algal taxa, in which *Mesostigma* is basal to charophyte green algae and sister to chlorophyte green algae [11].

We extracted 73 unique genes from the six genomes. Actual number of genes included in each genome ranges from 76 to 80 due to duplicated genes in IR. The gene set includes 62 characterized protein-coding genes, 3 rRNAs, 7 tRNAs (identified by amino acid anticodons) and a hypothetical conserved open reading frame (*ycf1*). The encoding reflects the order and orientation of genes in the genome. Location of multi-exon genes was determined by the starting position of the first exon. In one case,

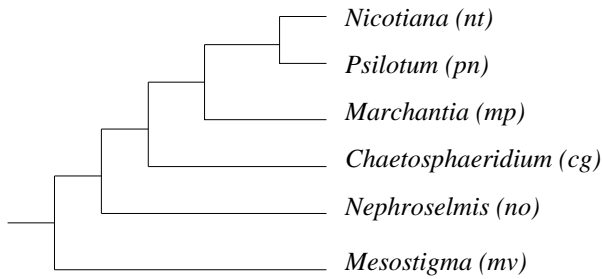


Fig. 1. The reference phylogeny of chloroplast genomes from land plants and green algae

the order of overlapping genes (*psbD-psbC*) was determined by the position of the start codon.

3. Mapping Contents in Ancestral Genomes

The first step of our new method—GRAPPA-IR is to determine the gene contents for each region in the genomes involved. Based on our previous research [32], such mapping can dramatically reduce the search space and improve the overall accuracy. When a genome is on a leaf (i.e., it is an extant taxon), we can easily determine the gene content for the LSC, SSC and IR regions through direct observation. However, since we do not know the gene order at each internal genome, we can only estimate the gene content for each region based on the assumption that all evolutionary events that alter the gene order are rare and that concurrent (i.e., parallel) changes in two children are less likely than a change in the parent. Thus, at each internal node, for a given region, when the regional gene contents for the two children are known, we face three possibilities of assigning a gene to the region:

- 1) If both children have gene *g* in the same region, then the parent has *g* in that region; otherwise, both children need to expand (or shrink) IRs and include *g* in that region, with a very low probability.
- 2) If neither child has *g*, then *g* is absent in the parent. Since the genomes we test all share 70 unique genes, we do not consider this case.
- 3) If *g* is located in different regions between the children, then it could be in either region of the parent. The two choices are equally likely without further information from the phylogeny. If the tree is rooted, we use the gene content in the evolutionary path to break the tie; otherwise, we are left with an undetermined outcome for *g*.

If a gene is undetermined in some internal node, it may become determined through a propagation of decisions from the leaves to the root. The estimated gene content for the internal nodes of the reference tree is presented in Fig. 2.

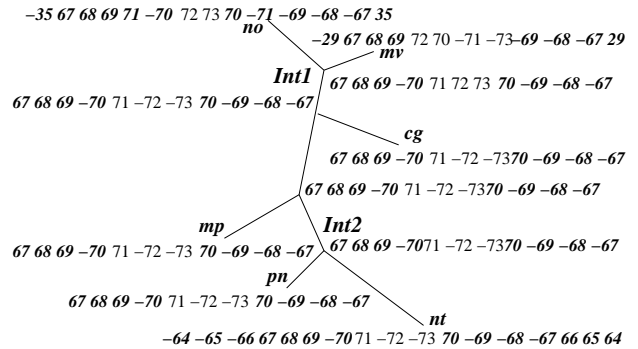


Fig. 2. Estimated gene contents for each region (only IR and SSC are shown).

Fig. 2 shows that the gene contents of the IR and SSC regions vary among the genomes. However the gene order of part of IR is highly conserved. For example, some genes (*rrn5*, *rrn16*, *rrn23*, coded as 67, 68 and 69, respectively) are always kept together in the IR. If we assume that inversions do not cross the IR boundary in most chloroplast genomes, then the evolution of chloroplast genome structure can be hypothesize as undergoing the following two steps:

- 1) A genome was divided into regions and inversions occurred independently in each region.
- 2) A segment from single copy regions was copied twice and joined to the existing inverted repeats, and the new genomes with longer IRs propagated.

One should notice that the above two steps could happen several times along each edge. Based on the above assumption, we can further simplify the gene content of IR and SSC, so that in the evolutionary path, IR regions for all genomes (leaves and internal) contain gene (67 68 69), and the SSC regions contain gene (70 71 72 73). This operation treats duplicate genes at the boundaries of IR and SSC as the last step towards the observed gene orders in the evolutionary path. Thus it is possible to ignore the duplications and reduce the problem to all leaf genomes of equal gene content (or with deletions). The simplified gene content map is shown in Fig. 3.

4. Ancestral Genome Reconstruction

We can reconstruct the phylogeny after the regional gene contents of all genomes are determined.

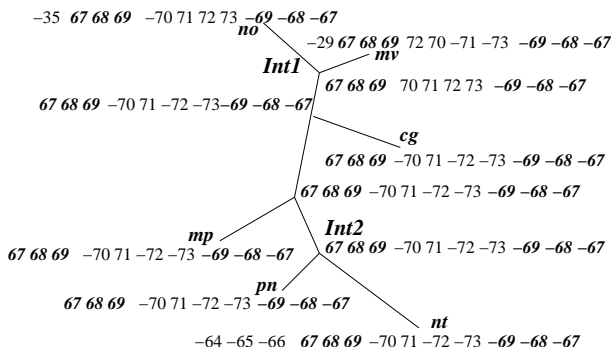


Fig. 3. Revised estimation of gene contents for each region (only IR and SSC are shown).

Since the gene contents are reduced to equal after the simplification step, it is feasible to use GRAPPA to infer an inversion phylogeny. That's to say, if inversions are allowed to cross the boundaries of pre-determined IR and single copy regions, we can use the original GRAPPA to compute the ancestral gene orders and the phylogeny. However, this is unlikely, since we do not observe any inversions involving genes in both the IR and single copy regions. Thus, we develop a new method called GRAPPA-IR that estimates inversions bounded by the boundary of IRs.

Bounding inversions into each region makes each region (*LSC*, *SSC* and *IR*) independent, which greatly simplify the computation. For two genomes G_1 and G_2 , the genomic distance between these two can be defined as $d(G_1, G_2) = d(SSC_1, SSC_2) + d(LSC_1, LSC_2) + d(IR_1, IR_2)$, i.e., the overall distance is the summation of all regional distances. If no gene content is changed for each region, then the regional distance can be computed using the linear algorithm proposed by [1]. However, if the regional content is not equal (deletions or insertions occur), then more complex algorithm [7] should be used.

For three given genomes g_1 , g_2 and g_3 , the median problem is to find a genome g_0 that can minimize the sum of distances from itself to the three given genomes. Although it is the simplest case for ancestral genome estimation, it is proved to be NP hard even for the simplest distance of breakpoint. Since we can deal with each region independently, the median problem can also be divided into three regional median problems, each of which is constructed from genes in the same region of genomes g_1 , g_2 and g_3 . After the regional median is obtained, the median solution on the whole genome can be constructed simply by concatenating the regions together. Again,

the regional median problems can be solved using any of the available inversion median solvers, such as [4] (for equal regional gene content) or [32] (with deletions/insertions).

To analyze dataset with more than three genomes, GRAPPA-IR uses an exhaustive approach devised for the original GRAPPA—it must test all possible trees to find the one with the minimum number of inversions. For each tree, the program tests a lower bound [17] to determine whether the tree is worth scoring; if so, then the program will iteratively solve the median problems at internal nodes and update the internal genomes, until no change occurs. GRAPPA-IR will not only return the best tree, but also estimated ancestral (internal) genomes with respect to that tree topology.

5. Experimental Results

5.1. Analysis of Six Chloroplast Genomes

We evaluate all trees for the six genomes using the new method. After 100 min of computation on a PIV 3.4GHz workstation, GRAPPA-IR returned a best tree with 74 inversions, with a topology agrees with the reference tree (Fig. 4).

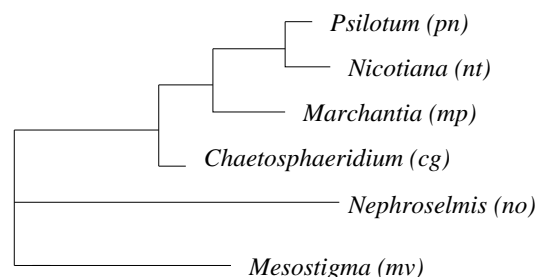


Fig. 4. The best tree obtained by GRAPPA-IR. The topology is the same as the reference tree.

We tested this data set with the original GRAPPA ignoring the region boundaries as well. The inference allows inversions to occur across the IR and single copy regions. The best tree obtained requires 73 inversions, yet the topology (Fig. 5) is very different from the reference trees and is in conflict with the biological relationship of these organisms. Although GRAPPA is a heuristic, extensive testing on simulated and biological data confirmed its high accuracy, thus its failure in this test suggests that the IR-boundary do perform a unique role in the evolution of chloroplast genomes, and a better method as GRAPPA-IR should be preferred.

We examined the reconstructed ancestral gene orders on the best tree returned by GRAPPA-IR. This

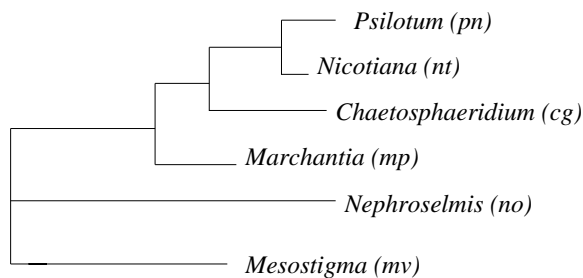


Fig. 5. The best trees when no IR boundary is imposed, with score 73. Notice that the topologies are different from the reference tree.

test dataset contains chloroplast genomes of land plants, green algae and a flagellate protist, which were separated by at least 450 million years of evolution [10]. The ancestral chloroplast genomes of land plants and algae contain inverted repeats, which is consistent with the hypothesis that IR is a feature derived early in the chloroplast endosymbiosis [21]. Although the sequenced cyanobacteria *Nostoc* and *Synechococcus* do not maintain *rrn*-containing IRs, if other cyanobacteria are identified with structures similar to the chloroplast IR, then it would suggest an even earlier origin for this structure. In addition, the ancestral IR contains the same gene content to that of *Mesostigma*, which agrees with the observation that *Mesostigma* chloroplast genome encode several ancestral gene clusters [14]. By comparison of ancestral gene orders to the extant genomes, it is possible to test formally the evolutionary force of gene order changes. For example, ancestral gene clusters may be more likely to be maintained if they share related function and are under constraints in the face of genome rearrangements.

5.2. Simulations to Assess Accuracy

Phylogenetic analysis methods deal with lost historic information, thus their accuracy should also be assessed through simulations, where the true evolutionary history is known. For this purpose, we generated datasets of 6 and 10 genomes, each with 78 genes (70 genes in LSC, 5 in SSC and 3 in IR), roughly in the range of our dataset described in the paper. We used a large range of evolutionary rates: let r denote the expected number of evolutionary events along an edge of the model tree, we used values of r in the range of 4 – 10. The actual number of inversions along each edge is sampled from a uniform distribution on the set $\{1, 2, \dots, 2r\}$. Given the model tree, we assigned the identity gene

order to the root, and randomly generated gene order for each node based on the edge length and the gene order of its parent, with the assumption that inversions can not cross the IR boundaries. For each combination of parameter settings, we simulated 20 datasets and averaged the results.

Given an inferred tree (reconstructed phylogeny), we can assess the topological accuracy in terms of *false positives* and *false negatives* [28] with respect to the true tree. If an edge in the true tree is missing in the inferred tree, this edge is then called a *false negative* (FN). Similarly, a *false positive* edge (FP) appears in the inferred tree, but not in the true tree. The FP and FN rates are the number of false positives (resp., false negatives) divided by the number of edges in the true tree. Since the inferred ancestral genomes have direct impact on the tree scores, thus we can examine the quality of these genomes simply by comparing the best tree score returned by GRAPPA-IR with the known true tree score.

We compared GRAPPA-IR to the original GRAPPA. We considered all trees with the minimum score given by both methods and took their strict consensus (of course, most time there is only one single best tree was returned). Therefore, the trees returned by both methods need not to be fully resolved and they tend to have somewhat better rates for false positives than for false negatives. Thus we report FN rates rather than FP rates or a single Robinson-Foulds score [28].

Figure 6 shows the results. This simulation indicates that GRAPPA-IR is clearly more accurate than the original GRAPPA for datasets with $r < 10$, which is in accordance with results on the six-genome dataset. Meanwhile, averagely speaking, scores of the best trees returned by GRAPPA-IR was only about 7% less than the true scores, which strongly indicates that the inferred ancestral gene orders should be very close to the true ancestral genomes.

6. Discussion

6.1. Mechanisms of IR Expansion

The gene content of the IR varies across land plants, even in a single genus or family [9]. It is known that homologous recombination is frequent between the two copies of IR [22]. In a single chloroplast, hundreds of copies of chloroplast DNA co-exist as circular monomer, dimer and linear chromosomes. In the cellular endosymbiosis environment, the selection on accuracy of replication may have

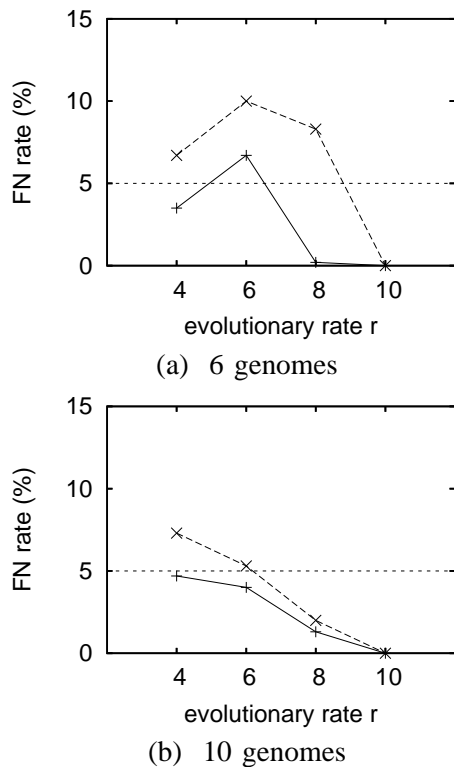


Fig. 6. FN rate for GRAPPA-IR (solid line) and GRAPPA (dashed line) as a function of the evolutionary rate r for 6 and 10 genomes. The horizontal line indicates the 5% error level, a typical threshold of acceptability for accurate phylogenetic reconstruction [30].

been relaxed to the degree that unequal recombination and replication slippage contribute to the expansion or shrinkage of IRs. Short repeat motifs may facilitate inter-molecular recombination and create diversity of chloroplast genomes in a population [12]. On the other hand, the intra-molecular recombination process should homogenize the sequences of the two IRs and thus the particular IR size and the gene content are maintained. The two counteracting phenomena may have played important roles in shaping the current diversity of chloroplast genome gene orders.

6.2. Duplications and Genome Stability

We found that incorrect gene order phylogenies were recovered without considering the IR boundary information. This suggests that maintenance of the IR is necessary in the evolution of chloroplast genomes in most of the cases. We propose that IR provides an insulation mechanism that stabilizes the genome structure, and the genes in single copy regions do not commute across the IR. This agrees with the observation that gene rearrangements are

more frequent in chloroplast genomes without IR [20]. However, some genomes with residual IRs but infrequent gene movements between single copy regions compared to related lineages do not conform to the hypothesis [5]. Future experimental studies on highly rearranged chloroplast genomes, for example, in the green alga *Chlamydomonas* lineage, may shed light on the maintenance of IR and genome rearrangements.

6.3. Comparison to Other Methods

A similar approach used for human and mouse genome comparison showed the optimal sorting of X chromosomes by at least 7 inversions [24]. This is a moderate amount of change compared to the level we observe in many chloroplast genomes. If duplications and deletions are considered in a finer scale, the process will be much complex, as suggested by the reconstruction of one 1.1 Mb region in the eutherian mammal ancestor [2]. Extensive tests show that trees returned by GRAPPA are superior to those returned by other gene-order phylogeny methods. The closely related package of Pevzner's group, MGR [3], is the only one that approaches its accuracy. All methods proposed in this paper could be applied to MGR and other method, which will result in a whole new set of tools for botanists interested in genome level evolution. A combination of DCM-GRAPPA and other approaches may scale up the capability to infer ancestral gene order for large genomes [31].

7. CONCLUSIONS

We implement a new method to infer ancestral gene orders with inverted repeats. Tests on a real dataset show accurate recovery of the genome phylogeny as well as inference of ancestral gene orders. This provides new insight into the genome evolutionary process. There still lie challenges to apply the method to large data sets of bacterial or eukaryotic genomes.

ACKNOWLEDGEMENTS

The work was supported by the US National Science Foundation grants DBI 0115684, DEB 0120709 to C.D. ANI 02-03584, EF 03-31654, IIS 01-13095, IIS 01-21377, DEB 01-20709 to B.M., and US National Institutes of Health grants R01 GM078991-01 to J.T.

References

- [1] Bader, D.A., Moret, B.M.E. and Yan, M. (2001). A linear-time algorithm for computing inversion distance between signed permutations with an experimental study. *J. Comput. Biol.*, **8**, 483–491.
- [2] Blanchette, M., Kunisawa, T. and Sankoff, D. (1999). Gene order breakpoint evidence in animal mitochondrial phylogeny. *J. Mol. Evol.*, **49**, 193–203.
- [3] Bourque, G. and Pevzner, P.A. (2002). Genome-Scale Evolution: Reconstructing Gene Orders in the Ancestral Species. *Genome Res.*, **12**, 26–36.
- [4] Caprara, A. (2001). On the practical solution of the reversal median problem. In *Proc. 1st Int'l Workshop Algs. in Bioinformatics (WABI'01)*, *Lecture Notes in Computer Science*, **2149**, 238–251.
- [5] Cosner, M.E., Raubeson, L.A. and Jansen, R.K. (2004). Chloroplast DNA rearrangements in Campanulaceae: phylogenetic utility of highly rearranged genomes. *BMC. Evol. Biol.*, **4**, 27.
- [6] Downie, S.R. and Palmer, J.D. (1992). Restriction Site Mapping of the Chloroplast DNA Inverted Repeat - a Molecular Phylogeny of the Asteridae. *Annals of the Missouri Botanical Garden*, **79**, 266–283.
- [7] El-Mabrouk, N. (2000). Genome rearrangement by reversals and insertions/deletions of contiguous segments In *Proc. 11th Ann. Symp. Combinatorial Pattern Matching (CPM'00)*, Volume 1848 of *Lecture Notes in Computer Science*, pp. 222–234. Springer-Verlag.
- [8] Gibbs, R.A., Weinstock, G.M., et al. (2004). Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature*, **428**, 493–521.
- [9] Goulding, S.E., Olmstead, R.G., Morden, C.W. and Wolfe, K.H. (1996). Ebb and flow of the chloroplast inverted repeat. *Mol. Gen. Genet.*, **252**, 195–206.
- [10] Herrmann, R.G., Maier, R.M., et al. (2003). Eukaryotic genome evolution: rearrangement and coevolution of compartmentalized genetic information. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **358**(1429): 87–97.
- [11] Karol, K.G., McCourt, R.M., Cimino, M.T. and Delwiche, C.F. (2001). The closest living relatives of land plants. *Science*, **294**, 2351–2353.
- [12] Kawata, M., Harada, T., et al. (1997). Short inverted repeats function as hotspots of intermolecular recombination giving rise to oligomers of deleted plastid DNAs (ptDNAs). *Curr. Genet.*, **31**, 179–184.
- [13] Kent, W.J., Baertsch, R., Hinrichs, A., et al. (2003). Evolution's cauldron: duplication, deletion, and rearrangement in the mouse and human genomes. *Proc. Natl. Acad. Sci. USA*, **100**, 11484–11489.
- [14] Lemieux, C., Otis, C. and Turmel, M. (2000). Ancestral ch. genome in *Mesostigma viride* reveals an early branch of green plant evolution. *Nature*, **403**, 649–652.
- [15] Martin, W., Stoebe, B., Goremykin, V., et al. (1998). Gene transfer to the nucleus and the evolution of chloroplasts. *Nature*, **393**, 162–165.
- [16] Maul, J.E., Lilly, J.W., Cui, L., et al. (2002). The *Chlamydomonas reinhardtii* plastid chromosome: islands of genes in a sea of repeats. *Plant Cell*, **14**, 2659–2679.
- [17] Moret, B.M.E., Tang, J., Wang, L.S. and Warnow, T. (2002). Steps toward accurate reconstructions of phylogenies from gene-order data. *J. Comp. Syst. Sci.*, **65**, 508–525.
- [18] Moret, B.M.E., Siepel, A.C., Tang, J. and Liu, T. (2002). Inversion medians outperform breakpoint medians in phylogeny reconstruction from gene-order data. In *2nd Int'l Workshop Algs. in Bioinformatics (WABI'02)*.
- [19] Muller, K., Borsch, T., et al. (2004). Evolution of carnivory in Lentibulariaceae and the Lamiales. *Plant Biol (Stuttg)* **6**(4): 477–90.
- [20] Palmer, J.D. and Thompson, W.F. (1982). Chloroplast DNA rearrangements are more frequent when a large inverted repeat sequence is lost. *Cell*, **29**, 537–550.
- [21] Palmer, J.D. (1985). Evolution of Chloroplast and Mitochondrial DNA in Plants and Algae. In *Molecular Evolutionary Genetics*. Edited by MacIntyre, R.J.. New York: Plenum Press; pp. 131–240.
- [22] Palmer, J.D. (1985). Comparative organization of chloroplast genomes. *Ann. Rev. Genet.* **19**, 325–354.
- [23] Perry, A.S., Brennan, S., Murphy, D.J., et al. (2002). Evolutionary re-organisation of a large operon in adzuki bean chloroplast DNA caused by inverted repeat movement. *DNA Res.*, **9**, 157–162.
- [24] Pevzner, P. and Tesler, G. (2003). Genome Rearrangements in Mammalian Evolution: Lessons From Human and Mouse Genomes. *Genome Res.*, **13**, 37–45.
- [25] Plunkett, G.M. and Downie, S.R. (2000). Expansion and Contraction of the cp Inverted Repeat in Apiaceae Subfamily Apioideae. *Syst. Bot.*, **25**, 648–667.
- [26] Price, R.A., Calie, P.J., Downie, S.R., et al. (1990). Chloroplast DNA variation in the Geraniaceae - a preliminary report. In *Proc. int Geraniaceae symp.*, Monville, South Africa, 1990. pp. 235–244.
- [27] Raven, J.A. and Allen, J.F. (2003). Genomics and chloroplast evolution: what did cyanobacteria do for plants? *Genome Biol.*, **4**, 209.
- [28] Robinson, D.R. and Foulds, L.R. (1981). Comparison of phylogenetic trees. *Math. Biosciences*, **53**, 131–147.
- [29] Rokas, A. and Holland, P.W.H. (2000). Rare genomic changes as a tool for phylogenetics. *Trends in Ecol. and Evol.*, **15**:454–459, 2000.
- [30] Swofford, D.L., Olson, G., Waddell, P. and Hillis, D.M. (1996). Phylogenetic inference. In Hillis, D.M., Moritz, M. and Mable, B. (eds), *Molecular Systematics*, 2nd ed., Sinauer Associates, Sunderland, pp. 407–514.
- [31] Tang, J. and Moret, B.M.E. (2003). Scaling up accurate phylogenetic reconstruction from gene-order data. *Bioinformatics*, **19** Suppl 1, i305–312.
- [32] Tang, J., Moret, B.M.E., Cui, L. and dePamphilis, C.W. (2004). Phylogenetic Reconstruction from Arbitrary Gene-Order Data. In *Proc. 4th IEEE Symp. on Bioinformatics and Bioeng. (BIBE'04)*. Taichung, Taiwan. May 2004, pp. 592–599.