Phylogenetic Reconstruction from Complete Gene Orders of Whole Genomes

K. M. SWENSON

School of Computer and Communication Sciences
Swiss Federal Institute of Technology (EPFL)
EPFL IC LCBB, INJ 232, Station 14
CH-1015 Lausanne, Switzerland
E-mail: krister.swenson@epfl.ch

W. ARNDT

Department of Computer Science & Engineering
University of South Carolina
Columbia, SC 29208, USA
E-mail: arndtw@cse.sc.edu

J. TANG

Department of Computer Science & Engineering
University of South Carolina
Columbia, SC 29208, USA
E-mail: jtang@engr.sc.edu

B. M. E. MORET

School of Computer and Communication Sciences
Swiss Federal Institute of Technology (EPFL)
EPFL IC LCBB, INJ 230, Station 14
CH-1015 Lausanne, Switzerland
and
Swiss Institute of Bioinformatics
E-mail: bernard.moret@epfl.ch

Rapidly increasing numbers of organisms have been completely sequenced and most of their genes identified; homologies among these genes are also getting established. It thus has become possible to represent whole genomes as ordered lists of gene identifiers and to study the evolution of these entities through computational means, in systematics as well as in comparative genomics. While dealing with rearrangements is nontrivial, the biggest stumbling block remains gene duplication and losses, leading to considerable difficulties in determining orthologs among gene families—all the more since orthology determination has a direct impact on the selection of rearrangements. None of the existing phylogenetic reconstruction methods that use gene orders is able to exploit the information present in complete gene families—most assume singleton families and equal gene content, limiting the evolutionary operations to rearrangements, while others make it so by eliminating nonshared genes and selecting one exemplar from each gene family. In this work, we leverage our past work on genomic distances, on tight bounding of parsimony scores through linear programming, and on divide-and-conquer methods for large-scale reconstruction to build the first computational approach to phylogenetic reconstruction from complete gene order data, taking into account not only rearrangements, but also duplication and

loss of genes. Our approach can handle multichromosomal data and gene families of arbitrary sizes and scale up to hundreds of genomes through the use of disk-covering methods. We present experimental results on simulated unichromosomal genomes in a range of sizes consistent with prokaryotes. Our results confirm that equalizing gene content, as done in existing phylogenetic tools, discards important phylogenetic information; in particular, our approach easily outperforms the most commonly referenced tool, MGR, often returning trees with less than one quarter of the errors found in the MGR trees.

Keywords: phylogenetic reconstruction; whole-genome data; genomic distance; gene inversion; gene duplication; gene loss

1. Introduction

Phylogenetic reconstruction has for many years been based on alignments of the sequences of one or more orthologous genes and proteins. The accumulation of full genome sequences enables one to use much richer data: one can use hundreds of genes to build a more detailed picture of organismal evolution[1–3] or one can be even more ambitious and use every gene present in the genomes. In the latter category are simple content-based approaches, where the presence or absence of genes from the global inventory are the informational characters;[4,5] as these approaches represent the data in the form of bit strings where each position is a character, they can make use of existing software packages for analysis. Obviously, however, a complete genome sequence contains much information besides the individual sequences of constituent genes or the presence or absence of these genes: the genome sequence identifies an ordering of these genes along the chromosomes, as well as a direction of transcription. Moreover, disruption of this ordering is a relatively rare occurrence—a "rare genomic event".[6] Thus changes in the ordering are valuable study tools in phylogenetics as well as comparative genomics.

Phylogenetic methods based on gene orders are still in their infancy—see the survey of Moret and Warnow:[7] the problems faced are mathematically and computationally much more challenging than in sequence-based reconstruction and the models not as well understood. These methods have been applied to simple data, such as organellar genomes across sets of taxa where the gene content is highly conserved (and where, of course, the number of genes is quite small, on the order of 40 genes for animal mitochondria and 120 genes for plant chloroplast).[8–12] As one attempts to scale such analyses to cellular organisms, several problems arise. One is simply a problem with the data: annotations of complete cellular genomes are still in various stages of completion, so that identifying homologous gene families with high accuracy is a challenge. Another is the highly variable gene content (just in bacteria, obligate endosymbionts may have under 1,000 genes, while free-living bacteria may have over 5,000). A third is the widespread occurrence of gene duplications and losses: gene families, while mostly containing a single gene, may contain up to 100 homologs in bacteria and over 1,000 homologs in eukaryotes. Finally, the difference in scale is a huge challenge given that most algorithms proposed for the task exhibit exponential growth in running time as a function of the size of the problem.

Only a few attempts to use gene orders for reconstructing the phylogeny of a group of cellular organisms have been made to date. The first few reduced the gene-order data (which forms a single phylogenetic character with an immense number of possible states)

to a collection of much simpler characters, such as the presence or absence of adjacent gene pairs,[13] an approach later broadened into formal encodings of gene orders used in parsimony analyses (see the survey of Wang et al.[14]); other approaches used phylogenetically informative clusters of genes.[15–17] More recently Belda et al.[18] used a variant of these approaches on a set of 30 γ-proteobacteria: they chose 247 specific orthologs present in all 30 bacteria, thereby both reducing the size of the problems and sidestepping the issue of gene families. Many papers have appeared on phylogenetic reconstruction from gene-order data when each gene order is a (signed) permutation of a reference set—for a recent survey, see Moret and Warnow[7]—, the two most notable ones being our own GRAPPA[19] and the multichromosomal tool MGR.[20] Finally, Blin *et al.*[21] went one step further on a subset of 13 of the aforementioned γ-proteobacteria, by using a local, pairwise restriction on gene content rather than a global one. None of these attempts to date has explicitly taken duplications and losses into account nor attempted to model them as evolutionary events. Bayesian MCMC methods, such as BADGER,[22] suffer from similar issues.

We earlier developed a measure of genomic distance that, given a pair of genomes, returns an estimate of the total number of evolutionary events under the iDLR (insertions, duplications, losses, and rearrangements) separating these two genomes.[23,24] (An alternative based on the closely related notion of common intervals recently appeared.[25]) Simulations results show very high accuracy up to a high threshold of saturation (where the estimated distance starts lagging behind the true number of events). Pairwise distances alone can be used as a basis for distance-based reconstruction, as was done for 13 γ-proteobacteria (the same that would later be used by Blin *et al.*[21]) in the MS thesis of Earnest-DeYoung,[26] who found that the reconstructed phylogeny differed from the reference one of Lerat *et al.*[2] by a single SPR event—that is, a single subtree was misplaced, as would also later be the case in the reconstruction of Blin *et al.*[21] This work served as a proof of concept, but used a number of *ad hoc* measures to keep the computational work down, such as identifying groups of genes that always formed a contiguous group and taking advantage of the reference phylogeny to establish an asymmetric cost for gene gains and losses.

In this paper, we combine our genomic distance[24] with our tight bounding based on linear programming (LP)[27] to produce the first phylogenetic reconstruction method that attempts to return a most parsimonious tree in terms of a palette of evolutionary events that include insertion, duplication, and loss of genes (or gene segments) as well as inversions, using the complete gene orders with full gene families and no prior known orthologies (as the orthologies will obviously depend on the returned tree). We provide experimental results comparing our new approach to reconstruction based on the genomic distances alone (using neighbor-joining), to reconstruction by our same tool, but from genomes reduced to equal gene contents, and to reconstruction, again on the basis of equalized gene contents, by the MGR server[20] and by neighbor-joining (NJ).

Our results indicate that computing under the iDLR model (i.e., using the full genomic gene ordering) regularly improves results over using equalized gene contents, often significantly so—errors are commonly reduced by a factor of 4 or more. They also indicate that the parsimonious trees returned by our LP-based procedure are as good as or better than those returned by neighbor-joining. Under parameter settings with relatively modest

numbers of events, the two exhibit similar accuracy, indicating that the iDLR distance estimates are both close to additivity and quite distinct from each other. These findings echo practice with sequence data, and, as with sequence data, we find that increased deviations from ultrametricity (in the form of widely different total amount of evolution on different paths from the root to the leaves) create situations where NJ does increasingly worse than our LP-based procedure—until the pairwise distances grow large enough to prevent accurate reconstruction by any means. We kept the number of taxa low (13 or fewer) in order to run large series of experiments with the LP-based method and with MGR, but we know from our past work[28] that the LP-based method can be scaled up to much larger numbers of genomes with very little loss of accuracy by using a disk-covering method.

2. Methods and Models

Our phylogenetic reconstruction algorithm is based on GRAPPA,[19,29] which we developed for analyzing chloroplast gene orders. GRAPPA examines every tree topology, computes a bound for each, and, for each tree that passes the bound, scores it by computing ancestral gene orders that minimize the total length of the tree, as measured in terms of inversions. The original GRAPPA is limited to singleton gene families and equal gene content, just like the various inference programs developed since, such as MGR, BADGER, etc. Its exhaustive examination of all trees also limits the maximum number of genomes it can handle, to about 15 taxa for single runs, 12-14 taxa when running benchmarks, while its method for scoring a tree requires the repeated computation of an inversion median at each internal node, an NP-hard problem that limits the lengths of tree edges it can handle. To extend it to larger numbers of taxa, Tang and Moret used a disk-covering method (in effect, a specialized divide-and-conquer approach) and showed that the resulting DCM-GRAPPA scaled gracefully to at least 1,000 genomes.[28] To date, the best way to extend the approach to larger genomes has been to avoid scoring trees. The original bounding computes a shortest cycle on the leaves of the tree and was found to eliminate well over 99% of the candidates.[29] Tang and Moret[27] later proposed a linear programming (LP) formulation where variables are the lengths of the tree edges and the constraints are simple metric inequalities; this approach eliminated well over 99.99% of the candidates in their experiments. Their LP formulation was later improved into a pure covering LP,[30] which offers efficient solutions (running in $O(n^{2.5})$ time, where n is the number of genes) and even more efficient approximations.

The LP score was close enough to the actual score that Tang and Moret proposed using the LP score in lieu of scoring the tree, avoiding any median computation. The resulting reconstruction lacks ancestral orderings, but gives a topology, an estimated score, and estimated edge lengths (the values of the LP variables), much as a maximum-likelihood reconstruction does for sequence data. We still lack a good approach to the inference of ancestral gene orders under the iDLR model, both from the point of view of computational effort (medians again) and from that of accuracy. Indeed, Earnest-DeYoung *et al.*,[31] in a study of the 13 γ-proteobacteria, found that internal gene orders were seriously underconstrained and so could not be reliably inferred—we need a more detailed and sensitive model of the evolutionary operations on a gene ordering.

The triangle inequalities that form the LP rely on a direct computation of the distances between selected pairs of leaves. Thus we can generalize the LP formulation directly to the iDLR model by using an estimate of the distance between two arbitrary genomes with varied gene families. We had proposed and tested just such a measure,[23,24] which estimates the total number of insertions (including duplications), losses, and inversions needed to transform one unichromosomal genome into another. The measure is readily extended to multichromosomal genomes by replacing inversions with double-cut-and-join operations,[32] since the latter cover fusion, fission, and translocation among chromosomes, yet can be handled just like inversions.

Our final algorithm thus combines DCMs for scaling to large numbers of genomes, a specific LP formulation to estimate branch lengths and total score of the trees, and the intergenomic distance of Swenson *et al.*[24] to provide input values to the LP. More specifically, we first compute the pairwise intergenomic distances; we then enumerate all possible trees, following the strategy of GRAPPA, attempting to eliminate as many trees as possible. The bounding is done first using the circular lower bound as described in Moret *et al.*;[33] if the tree passes that test, we then proceed to set up a linear program for it. In the linear program, the variables are the edge lengths; the constraints are derived using the triangle inequality—basically, a leaf-to-leaf path in the tree, corresponding to a particular sum of variables, should have length no less than the pairwise intergenomic distance between the two leaves. It should be noted that, whereas the constraints in the original use of the LP approach[27] were mathematically correct because all measures used were edit distances, the constraints used here have no such guarantee, since we are now using estimates of the true evolutionary distance. On the basis of the results of Swenson *et al.*,[24] we expect most of them to be correct, with a few possibly off by small deviations. Then again, we also expect the LP score to be even closer to optimal than in its original use, as the distances used in the constraints are much closer to the true evolutionary distances than was the case in the study of Tang and Moret. Finally, the score returned by the LP, rounded up to the nearest integer, is assigned as the score of that tree and the algorithm returns the trees with the lowest score.

3. Experimental Design

Our objective is to verify that computing under the full iDLR mode, i.e., handling both rearrangements and changes in gene content, allows for better reconstruction than handling only rearrangements on genomes reduced to signed permutations. Relative accuracy is thus our main evaluation criterion. However, absolute accuracy is needed in order to put the comparison in perspective. Since, in phylogenetic reconstruction, error rates larger than 10% are considered unacceptable, there is obviously little use in improving the error rate by a factor of two if the result is just bringing it from 60% down to 30%. We also need to test a wide range of parameters in the iDLR model, as well as to test the sensitivity of the methods to the rate of evolution. These considerations argue for testing on simulated data, where we can conduct both absolute and relative evaluations of accuracy, before we move to applying the tools to biological data, where only relative assessments of scores can be made. The range of dataset sizes need not be large, however, as we know that applying DCM methods

scales up results from datasets of fewer than 15 taxa to datasets of over one thousand taxa with little loss in accuracy and very little distortion over the range of parameters. As we can run many more tests on small datasets and as our primary interest is the effect of model parameters on accuracy, we generated datasets in the range of 10 to 13 taxa.

Simulated trees are often generated under the Yule-Harding model—they are birth-death trees. Many researchers observed that these trees are better balanced than most published ones. Other simulations have used trees chosen uniformly at random from the set of all tree topologies, so-called "random" trees; these, in contrast, are more imbalanced than most published trees. Aldous[34] proposed the β-split model to generate trees with a tailored level of balance; depending on the choice of β, this model can produce random trees ($\beta = -1.5$), birth-death trees ($\beta = 0$), and even perfectly balanced trees. We use all three types of trees in our experiments; for β-split trees, Aldous recommended using $\beta = -1$ to match the balance of most published trees; instead, we chose the parameter to match the computational effort on the datasets from which those trees were computed, which led us to using $\beta = -0.8$. On random and β-split trees, expected edge lengths are set after the tree generation by sampling from a uniform distribution on values in the set $\{1, 2, \ldots, r\}$, where r is a parameter that determines the overall rate of evolution. In the case of birth-death trees, we used both the same process and the edge lengths naturally generated by the birth-death process, deviated from ultrametricity and then scaled to fit the desired diameter.

We generate the true tree by turning each edge length into a corresponding number of iDLR evolutionary events on that edge. The events we consider under the iDLR model are insertions, duplications, losses, and inversions of genes or contiguous segments made of several genes—in particular, inserting, duplicating, or deleting a block of k consecutive genes has the same cost regardless of the value of k. We forced the expected number of inserted and duplicated elements to equal the expected number of deleted elements, in order to keep genome sizes within a general range. We varied the percentage of inversions as a function of the total number of operations from 20% to 90%. The remaining percentages were split evenly between insertions/duplications and losses, with the balance of insertions and duplications tested at one quarter, one half, and three quarters. The expected Gaussian-distributed length of each operation filled a range of combinations from 5 to 30 genes. These are conditions similar to, but broader in scope than, those used in the experiments reported in Swenson *et al.*[24]

In all our simulations, we used initial (root) genomes of 1'000 genes. The resulting leaf genomes are large enough to retain phylogenetic information while exhibiting large-scale changes in structure. These sizes correspond to the smaller bacterial genomes and allow us to conclude that our results will extend naturally to all unichromosomal bacterial genomes.

The collections of gene orders produced by these simulations are then fed to our various competing algorithms. These are of two types: (i) algorithms running on the full gene orders, namely NJ and our new LP-based algorithm; and (ii) algorithms running on equalized gene contents, which include NJ again (running on the inversion distance matrix produced by GRAPPA), GRAPPA, and MGR. Gene contents are equalized by removing gene families with more than one gene, then keeping only singleton genes common to all genomes. On some of these datasets, the equalized gene content is minuscule—with high rates of

evolution, the number of genes shared by all 12 taxa is occasionally in the single digits, obviously leading to serious inaccuracies on the part of reconstruction algorithms. We collect the data (including running times, the actual trees, internal inferred gene orders, inferred edge lengths, etc.) and compute basic measures, particularly the Robinson-Foulds[35] distance from the true tree—the most common error measure in phylogenetic reconstruction.

4. Results and Discussion

We ran collections of 100 datasets of 10 to 13 genomes, each of 1'000 genes, under various models of tree generation and various parameters of the iDLR model. We used birth-death, random, and β-split (with $\beta = -0.8$) models, with evolutionary diameters (the length of the longest path, as measured in terms of evolutionary operations, in the true tree) of 200, 400, 500, and 800 operations. (We ran tests with diameters of 800, but noted that most resulting instances exhibited strong saturation—that is, that many of the true edge lengths were significantly larger than the edit distances between the genomes at the ends of the edge; since no reconstruction method can do well in the presence of strong saturation, we did not pursue diameters larger than 800.) For each tree returned, we measured its RF error rate (the percent of edges in error with respect to the true tree) and then averaged the ratios over the set of test instances for each fixed parameter. We computed the ratio of the RF rate for our approach with that for NJ on full genomic distances and with those for the three approaches with equalized gene contents, binning the results into one "losing" bin (the other method did better), one bin of ties, and 5 bins of winners, according to the amount of improvement. Not all 100 instances are included in these averages, because some instances had equalized gene contents of just 2 or 3 genes and could not be run with GRAPPA.

We present below a few snapshots of our results. Table 1 shows the results of using full genomic distances for β-split trees on datasets of diameters 200, 400, and 500, using 80% inversions. In this case, no difference was found between the results returned by our LP-based method and those returned by NJ using full genomic distances. The average RF error rate for MGR was 23% for diameter 200, 32% for diameter 400, and 42% for diameter 500. As simple a method as NJ handily beats existing methods that must rely on equalized gene contents, often by large factors (e.g., factors of 4 or more in 26% of the cases with diameter 200 with respect to MGR). The reduction in error rate was sufficient in many cases to turn unacceptable results (with error rates well in excess of 10%) into acceptable ones.

Table 1. *Accuracy results for NJ on full genomic distances and for three evolutionary diameters compared to three methods on equalized gene contents. Column triples show wins, ties, and losses, in percent. Quintiles in the winning columns denote error reductions by factors larger than 4, 3, 2, 1.5, and 1.*

Dataset	NJ			GRAPPA			MGR		
200	16-4-25-1-0	50	4	14-0-11-4-0	1	3	26-6-21-4-1	36	6
400	4-0-5-4-0	23	0	3-0-6-1-0	0	0	5-1-7-6-12	1	4
500	5-5-5-8-0	69	8	11-2-14-17-15	18	23	17-7-14-17-14	24	7
	w	**t**	**l**	**w**	**t**	**l**	**w**	**t**	**l**

Experience with sequence data leads us to expect that an MP method, should do better than NJ when the diameter and deviation from ultrametricity get large. Our LP-based approach is a hybrid: unlike an MP method, it does not reconstruct ancestral labels, but like an MP method, it attempts to minimize the total length of the tree; thus it should at least occasionally outperform NJ. We tested this hypothesis on random trees and birth-death trees where, in both cases, we generated edge lengths by uniform sampling from the set $\{1, 2, \ldots, r\}$, for values of r ranging from 20 to 100, still using 80% inversions. Tables 2 and 3 present the results, this time limited to the reference MGR and to the two methods using full genomic data. Both tables show gains for the LP-based method over simple NJ

Table 2. *Error rates, in percent, on random trees for the two approaches using full genomic data and for MGR on equalized gene contents.*

	20	40	60	80	100
LP	0.9	8.0	7.8	6.0	26.0
NJ	0.5	8.5	8.7	9.5	25.5
MGR	11.3	31.8	34.0	35.0	49.0

Table 3. *Error rates, in percent, on birth-death trees for the two approaches using full genomic data and for MGR on equalized gene contents.*

	20	40	60	80	100
LP	0.2	8.5	7.6	5.7	19.4
NJ	1.4	9.0	8.5	8.0	18.0
MGR	9.7	31.7	31.8	33.7	51.4

as evolutionary rates increase, until both methods start failing at $r = 100$. Note that the accuracy gains over MGR are consistently very high.

Keeping the proportions of inversions to 80%, however, is neither very realistic, as gene duplications and losses are presumably more frequent in nature than rearrangements, nor very challenging, as, given a bounded set of possible gene choices, duplications and losses will saturate sooner than inversions. The experiments of Swenson *et al.*[24] did not test low percentages of inversions, so we ran sets of tests with 20% inversions only, keeping all other relative percentages of events identical. Table 4 shows these results. We were pleased, and somewhat surprised, to observe actual improvements in the quality of trees for rates up to $r = 40$; the threshold effect to $r = 60$ corresponds to a type of saturation caused by too many insertions and deletions. (Approaches with equalized gene contents are not reported, since they failed completely, as expected.)

Table 4. *Error rates, in percent, on birth-death trees with only 20% inversions.*

	20	40	60	80
LP	3.8	3.0	21.0	37.8
NJ	3.1	4.9	18.9	33.7

Finally, we reproduced the results of Earnest-DeYoung[26] on the dataset of 13 bacteria, with genome sizes ranging from 1'000 to over 5'000 genes and gene families of up to 70 members, this time without any special preprocessing, and using our LP-based approach rather than NJ. Once again the resulting phylogeny is one SPR (subtree) move away from that of Lerat et al. The large disparity in gene content between species in this dataset was handled automatically, for the first time for this dataset (or, indeed, for any other set of cellular genomes).

5. Conclusion

Our algorithm offers, for the first time, the possibility to evaluate the phylogenetic information present in the gene families and in the change in gene content among genomes while at the same time taking into account the complete gene orders; and they can do so on scales compatible with the smaller cellular genomes, such as bacterial genomes. Most importantly, our experiments indicate clearly the benefit to be derived from considering the full gene orderings of the genomes rather than some simplified subset—in almost all of our test cases, even the simple NJ procedure outperformed, often by large margins, the best reconstruction algorithms running on data with equalized gene contents. Much work remains to be done, of course: we need to generalize the distance computation of Swenson *et al.* to multichromosomal genomes (not particularly difficult using the DCJ model, but the introduction of additional parameters means further modelling questions) and to start using the algorithm on biological data, which should enable us to refine the model. And, while being able to estimate the true edge lengths of the tree is a help, we are still very far from being able to reconstruct ancestral genomes, because we have no viable algorithm to solve the vexing problem of the median of three genomes and because the iDLR model remains underconstrained.

Acknowledgments

JT and WA were supported by US National Institutes of Health (NIH) grant R01 GM078991-01 and by the University of South Carolina; WA also acknowledges support from the Rothberg Foundation.

References

1. J.R. Brown, C.J. Douady, M.J. Italia, W.E. Marshall and M.J. Stanhope, *Nature Genetics* **28**, 281 (2001).
2. E. Lerat, V. Daubin and N. Moran, *PLoS Biology* **1** (2003).
3. A. Rokas, B.L. Williams, N. King and S.B. Carroll, *Nature* **425**, 798 (2003).
4. S.T. Fitz-Gibbon and C.H. House, *Nucleic Acids Res.* **27**, 4218 (1999).
5. X. Gu and H. Zhang, *Mol. Biol. Evol.* **21**, 1401 (2004).
6. A. Rokas and P. Holland, *Trends in Ecol. and Evol.* **15**, 454 (2000).
7. B.M.E. Moret and T. Warnow, Advances in phylogeny reconstruction from gene order and content data, in *Molecular Evolution: Producing the Biochemical Data, Part B*, eds. E. Zimmer and E. Roalson, Methods in Enzymology **395**, 673 (Elsevier, 2005).
8. J.L. Boore and W.M. Brown, *Curr. Opinion Genet. Dev.* **8**, 668 (1998).
9. M.E. Cosner, R.K. Jansen, B.M.E. Moret, L.A. Raubeson, L.-S. Wang, T. Warnow and S.K. Wyman, An empirical comparison of phylogenetic methods on chloroplast gene order data

in Campanulaceae, in *Comparative Genomics*, eds. D. Sankoff and J.H. Nadeau, 99 (Kluwer Academic Publishers, 2000).
10. S.R. Downie and J.D. Palmer, Use of chloroplast DNA rearrangements in reconstructing plant phylogeny, in *Molecular Systematics of Plants*, eds. D. Soltis, P. Soltis and J.J. Doyle, 14 (Chapman and Hall, New York, 1992).
11. D. Sankoff, G. Leduc, N. Antoine, B. Paquin, B.F. Lang and R. Cedergren, *Proc. Nat'l Acad. Sci., USA* **89**, 6575 (1992).
12. D.B. Stein, D.S. Conant, M.E. Ahearn, E.T. Jordan, S.A. Kirch, M. Hasebe, K. Iwatsuki, M.K. Tan and J.A. Thomson, *Proc. Nat'l Acad. Sci., USA* **89**, 1856 (1992).
13. Y.I. Wolf, I.B. Rogozin, A.S. Kondrashov and E.V. Koonin, *BMC Evol. Biol.* **1** (2001).
14. L.-S. Wang, R.K. Jansen, B.M.E. Moret, L.A. Raubeson and T. Warnow, *J. Mol. Evol.* **63**, 473 (2006).
15. T. Kunisawa, *J. Theor. Biol.* **213**, 9 (2001).
16. T. Kunisawa, *J. Theor. Biol.* **222**, 495 (2003).
17. T. Kunisawa, *J. Theor. Biol.* **239**, 367 (2006).
18. E. Belda, A. Moya and F.J. Silva, *Mol. Biol. Evol.* **22**, 1456 (2005).
19. B.M.E. Moret, S.K. Wyman, D.A. Bader, T. Warnow and M. Yan, A new implementation and detailed study of breakpoint analysis, in *Proc. 6th Pacific Symp. on Biocomputing (PSB'01)*, 583 (World Scientific Pub., 2001).
20. G. Tesler, *J. Comp. Syst. Sci.* **65**, 587 (2002).
21. G. Blin, C. Chauve and G. Fertin, Genes order and phylogenetic reconstruction: Application to gamma-proteobacteria, in *Proc. 3rd RECOMB Workshop on Comparative Genomics (RECOMBCG'05)*, LNCS **3678**, 11 (Springer Verlag, 2005).
22. B. Larget, D.L. Simon, J.B. Kadane and D. Sweet, *Mol. Biol. Evol.* **22**, 486 (2005).
23. M. Marron, K.M. Swenson and B.M.E. Moret, *Theor. Comp. Sci.* **325**, 347 (2004).
24. K.M. Swenson, M. Marron, J.V. Earnest-DeYoung and B.M.E. Moret, Approximating the true evolutionary distance between two genomes, in *Proc. 7th SIAM Workshop on Algorithm Engineering & Experiments (ALENEX'05)*, 121 (SIAM Press, Philadelphia, 2005).
25. S. Angibaud, G. Fertin, I. Rusu and S. Vialette, *J. Comp. Biol.* **14**, 379 (2007).
26. J. Earnest-DeYoung, Reversing gene erosion: Reconstructing ancestral bacterial gene orders, Master's thesis, University of New Mexico (2004).
27. J. Tang and B.M.E. Moret, Linear programming for phylogenetic reconstruction based on gene rearrangements, in *Proc. 16th Ann. Symp. Combin. Pattern Matching (CPM'05)*, LNCS **3537**, 406 (Springer Verlag, 2005).
28. J. Tang and B.M.E. Moret, Scaling up accurate phylogenetic reconstruction from gene-order data, in *Proc. 11th Int'l Conf. on Intelligent Systems for Mol. Biol. (ISMB'03)*, Bioinformatics **19**, i305 (2003).
29. B.M.E. Moret, J. Tang, L.-S. Wang and T. Warnow, *J. Comp. Syst. Sci.* **65**, 508 (2002).
30. A. Bachrach, K. Chen, C. Harrelson, R. Mihaescu, S. Rao and A. Shah, Lower bounds for maximum parsimony with gene order data, in *Proc. 3rd RECOMB Workshop on Comparative Genomics (RECOMBCG'05)*, LNCS **3678**, 1 (Springer Verlag, 2005).
31. J.V. Earnest-DeYoung, E. Lerat and B.M.E. Moret, Reversing gene erosion: reconstructing ancestral bacterial genomes from gene-content and gene-order data, in *Proc. 4th Int'l Workshop Algs. in Bioinformatics (WABI'04)*, LNCS **3240**, 1 (Springer Verlag, 2004).
32. S. Yancopoulos, O. Attie and R. Friedberg, *Bioinformatics* **21**, 3340 (2005).
33. B.M.E. Moret, L.-S. Wang, T. Warnow and S. Wyman, New approaches for reconstructing phylogenies from gene-order data, in *Proc. 9th Int'l Conf. on Intelligent Systems for Mol. Biol. (ISMB'01)*, Bioinformatics **17**, S165 (2001).
34. D. Aldous, *Stat. Sci.* **16**, 23 (2001).
35. D. Robinson and L. Foulds, *Math. Biosci.* **53**, 131 (1981).