BMC Bioinformatics

SOFTWARE Open Access

MLGO: phylogeny reconstruction and ancestral inference from gene-order data

Fei Hu[1,2], Yu Lin[3] and Jijun Tang[1,2*]

Abstract

Background: The rapid accumulation of whole-genome data has renewed interest in the study of using gene-order data for phylogenetic analyses and ancestral reconstruction. Current software and web servers typically do not support duplication and loss events along with rearrangements.

Results: MLGO (Maximum Likelihood for Gene-Order Analysis) is a web tool for the reconstruction of phylogeny and/or ancestral genomes from gene-order data. MLGO is based on likelihood computation and shows advantages over existing methods in terms of accuracy, scalability and flexibility.

Conclusions: To the best of our knowledge, it is the first web tool for analysis of large-scale genomic changes including not only rearrangements but also gene insertions, deletions and duplications. The web tool is available from http://www.geneorder.org/server.php.

Keywords: Phylogeny reconstruction, Ancestral inference, Genome rearrangement, Maximum likelihood

Background

As whole genomes are sequenced at increasing rates, using gene-order data[a] for phylogenetic analyses and ancestral reconstruction is attracting increasing interest. Comparative genomics, evolutionary biology, and cancer research all require tools to elucidate the history and consequences of the large-scale genomic changes, such as rearrangements, duplications, losses. However, using gene-order data has proved far more challenging than using sequence data and numerous problems plague existing methods: oversimplified models, poor accuracy, poor scaling, lack of robustness, lack of statistical assessment, etc.

Genome rearrangement operations change the ordering of genes on chromosomes. An *inversion* operation (also called *reversal*) reverses both the order and orientation of a segment of a chromosome. A *transposition* is an operation that swaps two adjacent segments of a chromosome. In case of multiple chromosomes, a *translocation* breaks a chromosome and reattaches a part to another chromosome, while a *fusion* joins two chromosomes and a *fission* splits one chromosome into two. Yancopoulos *et al.* [1] proposed a universal *double-cut-and-join* (DCJ) operation that accounts for all rearrangements used to date. None of these operations alter the gene content of genomes, whereas *deletions* (or *losses*) delete segments of (one or more) contiguous genes from a chromosome, while *insertions* introduce a segment of (one or more) contiguous genes from external sources into a chromosome. and *duplications* copies an existing segment within the genome and inserts into a chromosome. Finally, *whole genome duplication* (WGD) creates an additional copy of the entire genome of a species.

As phylogenies play a central role in biological research, over the past decade many methods were developed to reconstruct phylogenies from gene-order data. The first algorithm for phylogeny inference from gene-order data was BPAnalysis based on breakpoint distances [2]. Moret *et al.* [3] later extended this approach with GRAPPA by using inversion distances. While these methods were limited to unichromosomal genomes, Bourque and Pevzner [4] developed MGR to handle multichromosomal genomes. These approaches are parsimony-based: they solve the so-called Big Parsimony Problem (BPP) and all suffer from serious scalability issues. In contrast

*Correspondence: jtang@ces.sc.edu
[1] Tianjin Key Laboratory of Cognitive Computing and Application, Tianjin University, 300072 Tianjin, China
[2] Department of Computer Science and Engineering, University of South Carolina, SC 29208 Columbia, USA
Full list of author information is available at the end of the article

with parsimony-based methods, distance-based methods run in time polynomial in the number and size of genomes. Lin *et al.* [5] have demonstrated the accuracy and scalability of a distance-based method that uses NJ [6] and FastME [7] with an accurate distance estimator [8]. Instead of working directly with the evolutionary events of the model, one can also transform the problem into the familiar sequence-based reconstruction problem. Wang *et al.* [9] first proposed a parsimony-based approach, MPBE (Maximum Parsimony on Binary Encoding). Recently Hu *et al.* [10] developed MLBE, later refined by Lin *et al.* [11] with MLWD, both of which demonstrate that using maximum-likelihood approaches is the decisive factor in improving the modest accuracy of MPBE.

If the tree is fixed, then computing its parsimony score is known as the Small Parsimony Problem (SPP). Ancestral reconstruction has been studied through several optimization schemes for SPP on gene-order data—using adjacencies [12-15], using conserved intervals (Roci—Reconstruction of Conserved Intervals [16]), using multiple breakpoint graphs (MGRA [17]) and supporting whole-genome duplications [18,19], where continuous regions or complete ancestral genomes have been inferred.

Relatively few of these tools are offered through web servers. Lin *et al.* [20] had developed a web-server version of MGR with new heuristics to speed up the original MGR algorithm, but the site is no longer accessible. Both Roci and MGRA (for ancestral reconstruction only) are offered through web servers, but none can handle complex events such as gene insertions, deletions and duplications.

We present a new tool MLGO for the reconstruction of phylogeny and/or ancestral genomes from gene-order data. MLGO relies on two methods we have developed: MLWD [11] for phylogenetic reconstruction and PMAG+ [21] for ancestral genome reconstruction. Our tool takes the advantage of binary encoding on gene-order data, supports a fairly general model of genomic evolution (rearrangements plus duplications, insertions, and losses of genomic regions), and successfully accommodates itself into the framework of maximized likelihood. The results of extensive testing on both simulated and real data show that both MLWD and PMAG+ can achieve great performance, scalability and flexibility, suggesting MLGO a suitable tool for large-scale analysis of high-resolution data. Furthermore, MLGO is deployed as a web service, providing the first web tool that is suitable for large scale genomic analysis with a general model of evolution.

Implementation

MLGO preprocesses the gene-order data, configures the transition model, reconstructs a phylogeny, and finally solves the SPP on that phylogeny.

Terminology

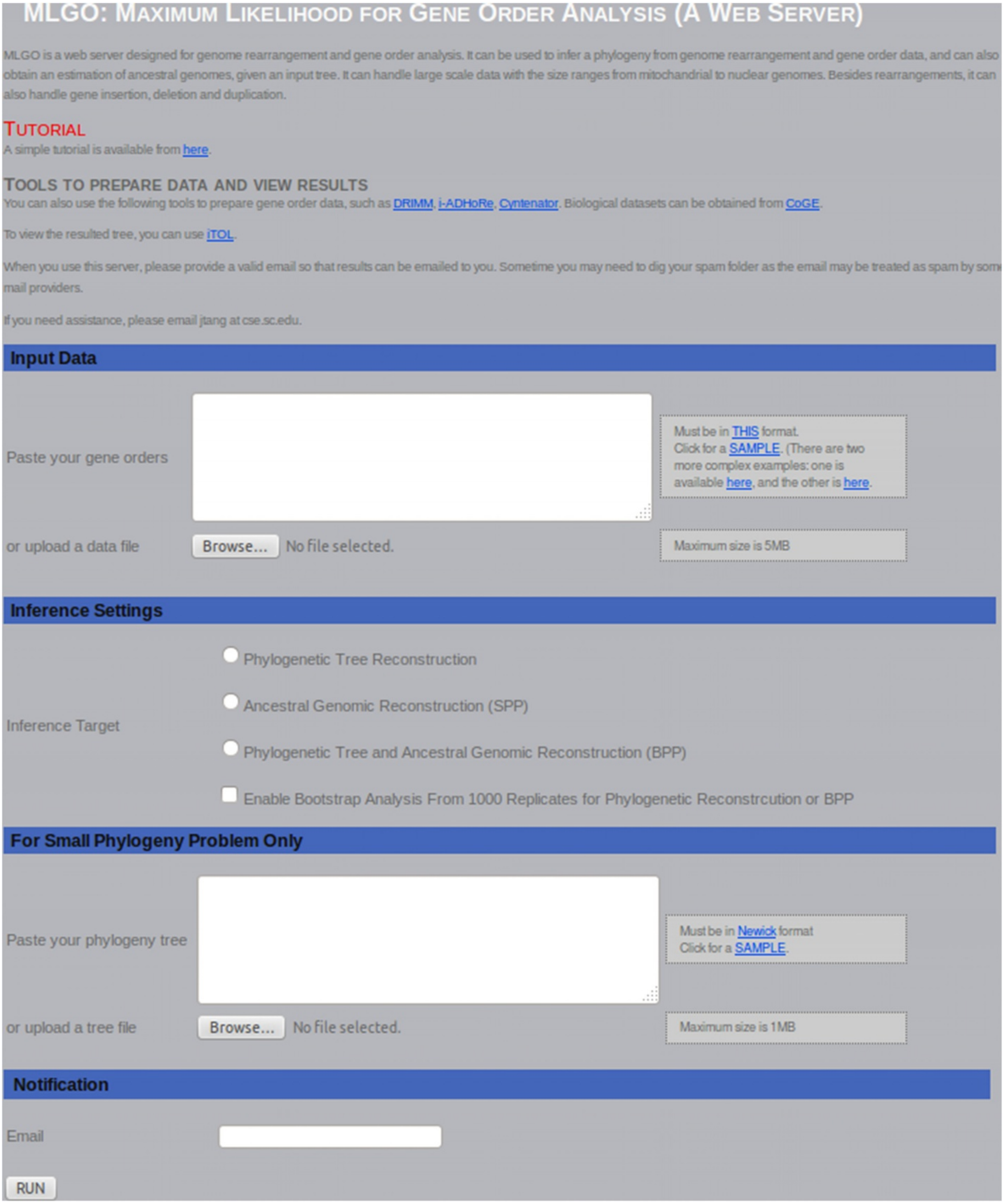
Given a set of n genes labeled as $\{1, 2, \cdots, n\}$, gene-order data for a genome consists of lists of genes in the order in which they are placed along one or more chromosomes. Each gene is assigned with an orientation that is either positive, written i, or negative, written $-i$. Two genes i and j form an *adjacency* (i, j) if i is immediately followed by j, or, equivalently, $-j$ is immediately followed by $-i$. If gene k lies at one end of a linear chromosome, we let k be adjacent to an extremity o to mark the beginning or ending of the chromosome, written as (o, k) or (k, o), and called *telomere*.

Phylogeny reconstruction

The data preprocessing and the configuration of the transition model follow the approach of MLWD [11]. Each adjacency that appears at least once in the collection of input genomes corresponds to a unique character position in the sequence and the presence or absence of any of these adjacencies in a given genomes is coded by a 1 (presence) or a 0 (absence). Since our encodings are binary sequences, the parameters of the model are simply the transition probability from presence (1) to absence (0) and that from absence (0) to presence (1). Lin *et al.* [11] gave the following derivation for these parameters. A DCJ operation selects uniformly at random two adjacencies (or telomeres) and replaces them by two new adjacencies (or telomeres). Since a genome with n genes and $O(1)$ chromosomes has $n + O(1)$ adjacencies and telomeres, the transition probability from 1 to 0 is $\frac{2}{n+O(1)}$ under one DCJ operation; and since there are up to $\binom{2n+2}{2}$ possible adjacencies and telomeres, the transition probability from 0 to 1 is $\frac{2}{2n^2+O(n)}$. Thus the transition from 0 to 1 is roughly $2n$ times less likely than that from 1 to 0. Despite the restrictive assumption that all DCJ operations are equally likely, this result is in line with the observed bias in transitions of adjacencies given by Sankoff and Blanchette [22]: the probability of breaking a given ancestral adjacency is high while that of creating a particular adjacency along several lineages is low (a version of homoplasy for adjacencies). Finally, the encoding adds characters and a transition probability for the presence or absence of each unique gene. Due to duplicated genes, there is no one-to-one correspondence between genomes and the final encodings of multisets of genes, adjacencies, and telomeres. Once we have the binary sequences and transition parameters, we can reconstruct a phylogeny using maximum likelihood. Of the many implementations of this method, we chose RAxML [23] for its speed and its dedicated handling of binary sequences.

Bootstrap support

A distinct advantage of using sequence encoding is the ability to use the bootstrap method to assess the



MLGO: MAXIMUM LIKELIHOOD FOR GENE ORDER ANALYSIS (A WEB SERVER)

MLGO is a web server designed for genome rearrangement and gene order analysis. It can be used to infer a phylogeny from genome rearrangement and gene order data, and can also obtain an estimation of ancestral genomes, given an input tree. It can handle large scale data with the size ranges from mitochandrial to nuclear genomes. Besides rearrangements, it can also handle gene insertion, deletion and duplication.

TUTORIAL

A simple tutorial is available from here.

TOOLS TO PREPARE DATA AND VIEW RESULTS

You can also use the following tools to prepare gene order data, such as DRIMM, i-ADHoRe, Cyntenator. Biological datasets can be obtained from CoGE.

To view the resulted tree, you can use iTOL.

When you use this server, please provide a valid email so that results can be emailed to you. Sometime you may need to dig your spam folder as the email may be treated as spam by som

mail providers.

If you need assistance, please email jtang at cse.sc.edu.

Input Data

Paste your gene orders

Must be in THIS format.
Click for a SAMPLE. (There are two more complex examples: one is available here, and the other is here.

or upload a data file

No file selected.

Maximum size is 5MB

Inference Settings

Inference Target

○ Phylogenetic Tree Reconstruction

○ Ancestral Genomic Reconstruction (SPP)

○ Phylogenetic Tree and Ancestral Genomic Reconstruction (BPP)

□ Enable Bootstrap Analysis From 1000 Replicates for Phylogenetic Reconstrcution or BPP

For Small Phylogeny Problem Only

Paste your phylogeny tree

Must be in Newick format
Click for a SAMPLE.

or upload a tree file

No file selected.

Maximum size is 1MB

Notification

Email

Figure 1 The screen shot of the web interface for MLGO.

robustness of the inferred phylogeny. Doing so with gene-order data is not possible, because a chromosome with n distinct genes presents a single character (the ordering) with $2^n \times n!$ possible states (the first term is for the strandedness of each gene and the second for the possible permutations in the ordering). This single character

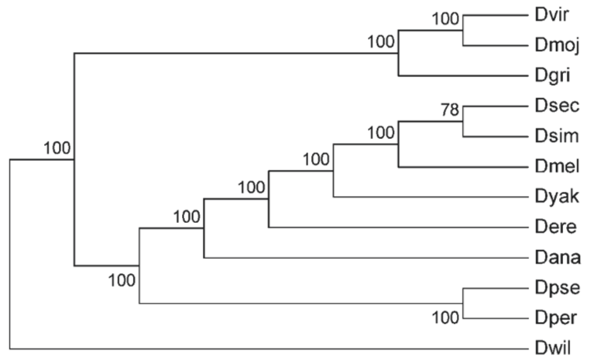


Figure 2 The consensus phylogeny of 12 drosophila genomes with bootstrap support values from 100 replicates.

is equivalent to an alignment with a single column, albeit one where each character can take any of a huge number of states—we cannot meaningfully resample a single character. The binary encoding effectively maps this single character into a high-dimensional binary vector, so that the standard phylogenetic bootstrap [24] can be used. While the evolution of a specific adjacency depends directly on several others, independence can be assumed if, once an adjacency is broken during evolution, it is not formed again—an analog of Dollo parsimony, but one that is very likely in rearrangement data due to the enormous state space [25].

Ancestral inference

Using the phylogeny thus computed, we then proceed to solve the SPP, now following the approach of Hu *et al.* [21]. The first step involves the estimation of ancestral gene contents from the contents of the input genomes. Our inference of ancestral contents relies on viewing genes and adjacencies as independent binary characters, as described for the encoding. Whether or not an ancestral genome contains a gene or an adjacency is determined by the conditional probability of the presence state of the gene or the adjacency, computed by the marginal probabilistic reconstruction method suggested by Yang *et al.* [26]. If such probability is larger than 50%, we conclude that the gene belongs to the genome. We extend this approach to compute the probability of observing each adjacency. We then reduce the adjacency assembly problem for any given ancestral genome to an instance of the Travelling Salesperson Problem (TSP), by representing genes as vertices and adjacencies as edges, and finally solve the TSP by using `Concorde` [27].

Results and discussion

`MLGO` is written in C++ and Perl as a web tool. Figure 1 shows the screen shot of the web interface for `MLGO`. The input format of the dataset is that used by `GRAPPA` and `MGR`: FASTA-like headers for the names of the genomes (> followed by an alphanumeric sequence followed by a newline), each chromosome represented by a signed permutation of integers ending with a $ symbol and a newline character. Phylogenies are output as trees in Newick format.

We used the genomes of 12 fully sequenced drosophila species to demonstrate the performance of `MLGO`. Figure 2 shows the consensus phylogeny reconstructed by `MLGO` with the bootstrap support values obtained using 100 replicates. Compared to the study using sequence data published by Clark *et al.* [28], all major groups in those 12 drosophila genomes were correctly identified with strong support (bootstrap value $>$ 90), except for one median support at the bipartition between *D. simulans*, *D. sechellia* and the rest. The total running time for reconstructing the phylogeny of 12 drosophila species is less than 1 minute, while ancestral reconstruction adds less than 30 minutes. We also tested the performance of `MLGO` on 15 Metazoan genomes from the eGOB (Eukaryotic Gene Order Browser) database [29], and the reconstructed phylogeny tree shown in Figure 3 is perfectly supported from existing studies [30,31].

Conclusion

As whole genomes are sequenced at increasing rates, using gene-order data for phylogenetic analyses and ancestral reconstruction is attracting increasing interest,

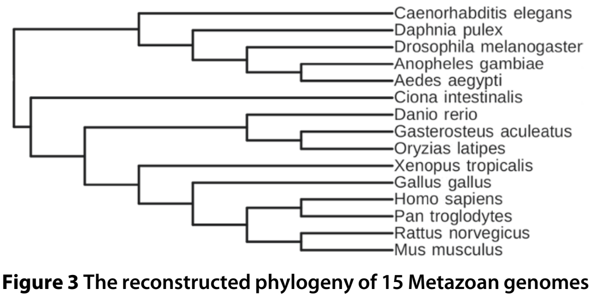


Figure 3 The reconstructed phylogeny of 15 Metazoan genomes.

especially coupled with the recent advances in identifying conserved synteny blocks among multiple species [32-34].

MLGO (Maximum Likelihood for Gene-Order Analysis) is the first web tool for likelihood-based inference of both the phylogeny and ancestral genomes. It provides fast and scalable analyses with bootstrap support of large-scale genomic changes including not only rearrangements but also gene insertions, deletions and duplications.

Availability and requirements

The web tool is available from http://www.geneorder.org/server.php.

Project name: MLGO
Project home page: http://www.geneorder.org/server.php
Operating system(s): Platform independent
Programming language: Perl
Other requirements: None
License: GNU
Restrictions for use by non-academics: None

Endnote

[a] We use the term "gene" as this is in fact a common form of syntenic blocks, but other kinds of markers could be used.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

FH implemented the web server. YL contributed to the phylogeny reconstruction part with the help of FH and JT. FH and JT contributed to the ancestral inference part. JT provided advice and oversight of the project. All authors drafted, read and approved the final manuscript.

Acknowledgements

We thank Bernard Moret for helpful discussions. FH and JT were funded by NSF IIS 1161586 and an internal grant from Tianjin University, China. YL was supported by a fellowship of the Swiss National Science Foundation (grant no. 146708). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Author details

[1]Tianjin Key Laboratory of Cognitive Computing and Application, Tianjin University, 300072 Tianjin, China. [2]Department of Computer Science and Engineering, University of South Carolina, SC 29208 Columbia, USA. [3]Department of Computer Science and Engineering, University of California, San Diego, CA 92093 La Jolla, USA.

Received: 16 July 2014 Accepted: 16 October 2014
Published online: 08 November 2014

References

1. Yancopoulos S, Attie O, Friedberg R: **Efficient sorting of genomic permutations by translocation, inversion and block interchange.** *Bioinformatics* 2005, **21**(16):3340–3346.
2. Blanchette M, Bourque G, Sankoff D: **Breakpoint phylogenies.** *Genome Inform* 1997, **1997:**25–34.
3. Moret B, Wang L, Warnow T, Wyman S: **New approaches for reconstructing phylogenies from gene order data.** *Bioinformatics* 2001, **17**(suppl 1):165–173.
4. Bourque G, Pevzner P: **Genome-scale evolution: reconstructing gene orders in the ancestral species.** *Genome Res* 2002, **12**(1):26–36.
5. Lin Y, Rajan V, Moret BME: **TIBA: a tool for phylogeny inference from rearrangement data with bootstrap analysis.** *Bioinformatics* 2012, **28**(24):3324–3325.
6. Saitou N, Nei M: **The neighbor-joining method: a new method for reconstructing phylogenetic trees.** *Mol Biol Evol* 1987, **4**(4):406–425.
7. Desper R, Gascuel O: **Fast and accurate phylogeny reconstruction algorithms based on the minimum-evolution principle.** *J Comput Biol* 2002, **9**(5):687–705.
8. Lin Y, Moret BME: **Estimating true evolutionary distances under the DCJ model.** *Bioinformatics* 2008, **24**(13):i114–i122.
9. Wang L-S, Jansen R, Moret BME, Raubeson L, Warnow T: **Fast phylogenetic methods for the analysis of genome rearrangement data: an empirical study.** In *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing (PSB)*. Singapore: World Scientific; 2001:524–535.
10. Hu F, Gao N, Zhang M, Tang J: **Maximum likelihood phylogenetic reconstruction using gene order encodings.** In *Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), 2011 IEEE Symposium On*. USA: IEEE; 2011:1–6.
11. Lin Y, Hu F, Tang J, Moret BME: **Maximum likelihood phylogenetic reconstruction from high-resolution whole-genome data and a tree of 68 eukaryotes.** In *Proc. 18th Pacific Symp. on Biocomputing, (PSB)*. Singapore: World Scientific; 2013:285–296.
12. Ma J, Zhang L, Suh BB, Raney BJ, Burhans RC, Kent WJ, Blanchette M, Haussler D, Miller W: **Reconstructing contiguous regions of an ancestral genome.** *Genome Res* 2006, **16**(12):1557–1565.
13. Ma J, Ratan A, Raney BJ, Suh BB, Zhang L, Miller W, Haussler D: **Dupcar: reconstructing contiguous ancestral regions with duplications.** *J Comput Biol* 2008, **15**(8):1007–1027.
14. Ma J: **A probabilistic framework for inferring ancestral genomic orders.** In *Bioinformatics and Biomedicine (BIBM), 2010 IEEE International Conference On*. USA: IEEE; 2010:179–184.
15. Gagnon Y, Blanchette M, El-Mabrouk N: **A flexible ancestral genome reconstruction method based on gapped adjacencies.** *BMC Bioinformatics* 2012, **13**(Suppl 19):4.
16. Bergeron A, Blanchette M, Chateau A, Chauve C: **Reconstructing ancestral gene orders using conserved intervals.** In *Proc. 4th Int'l Workshop Algs. in Bioinformatics (WABI'04)*. Germany: Springer; 2004:14–25.
17. Alekseyev MA, Pevzner PA: **Breakpoint graphs and ancestral genome reconstructions.** *Genome Res* 2009, **19**(5):943–957.
18. Murat F, Xu J-H, Tannier E, Abrouk M, Guilhot N, Pont C, Messing J, Salse J: **Ancestral grass karyotype reconstruction unravels new mechanisms of genome shuffling as a source of plant evolution.** *Genome Res* 2010, **20**(11):1545–1557.
19. Ouangraoua A, Tannier E, Chauve C: **Reconstructing the architecture of the ancestral amniote genome.** *Bioinformatics* 2011, **27**(19):2664–2671.
20. Lin CH, Zhao H, Lowcay SH, Shahab A, Bourque G: **webmgr: an online tool for the multiple genome rearrangement problem.** *Bioinformatics* 2010, **26**(3):408–410.
21. Hu F, Zhou J, Zhou L, Tang J: **Probabilistic reconstruction of ancestral genomes with gene insertions and deletions.** *IEEE/ACM Trans Comput Biol Bioinformatics* 2014, **11**(4):667–672.
22. Sankoff D, Blanchette M: **Probability models for genome rearrangement and linear invariants for phylogenetic inference.** In *Proc. 3rd Int'l Conf. Comput. Mol. Biol. (RECOMB'99)*. USA: ACM; 1999:302–309.
23. Stamatakis A: **Raxml-vi-hpc: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models.** *Bioinformatics* 2006, **22**(21):2688–2690.
24. Felsenstein J: **Confidence limits on phylogenies: an approach using the bootstrap.** *Evol* 1985, **39:**783–791.
25. Lin Y, Rajan V, Moret BME: **Bootstrapping phylogenies inferred from rearrangement data.** In *Proc. 11th Workshop Algs. in Bioinf. (WABI'11), Lecture Notes in Computer Science, Vol. 6833*. Germany: Springer; 2011:175–187.
26. Yang Z, Kumar S, Nei M: **A new method of inference of ancestral nucleotide and amino acid sequences.** *Genetics* 1995, **141**(4):1641–1650.
27. Applegate D, Bixby R, Chvatal V, Cook W: **Concorde tsp solver.** 2006. [http://www.tsp.gatech.edu/concorde]
28. Clark AG, Eisen MB, Smith DR, Bergman CM, Oliver B, Markow TA, Kaufman TC, Kellis M, Gelbart W, Iyer VN, Pollard DA, Sackton TB, Larracuente AM,

Singh ND, Abad JP, Abt DN, Adryan B, Aguade M, Akashi H, Anderson WW, Aquadro CF, Ardell DH, Arguello R, Artieri CG, Barbash DA, Barker D, Barsanti P, Batterham P, Batzoglou S, Begun D, et al.: **Evolution of genes and genomes on the drosophila phylogeny.** *Nature* 2007, **450**(7167):203–218.

29. López MD, Samuelsson T: **eGOB: eukaryotic gene order browser.** *Bioinformatics* 2011, **27**(8):1150–1151.
30. Ponting CP: **The functional repertoires of metazoan genomes.** *Nat Rev Genet* 2008, **9**(9):689–698.
31. Srivastava M, Begovic E, Chapman J, Putnam NH, Hellsten U, Kawashima T, Kuo A, Mitros T, Salamov A, Carpenter ML, Signorovitch AY, Moreno MA, Kamm K, Grimwood J, Schmutz J, Shapiro H, Grigoriev IV, Buss LW, Schierwater B, Dellaporta SL, Rokhsar DS: **The trichoplax genome and the nature of placozoans.** *Nature* 2008, **454**(7207):955–960.
32. Simillion C, Janssens K, Sterck L, Van de Peer Y: **i-adhore 2.0: an improved tool to detect degenerated genomic homology using genomic profiles.** *Bioinformatics* 2008, **24**(1):127–128.
33. Pham SK, Pevzner PA: **Drimm-synteny: decomposing genomes into evolutionary conserved segments.** *Bioinformatics* 2010, **26**(20):2509–2516.
34. Rödelsperger C, Dieterich C: **Cyntenator: progressive gene order alignment of 17 vertebrate genomes.** *PloS one* 2010, **5**(1):8861.

doi:10.1186/s12859-014-0354-6
Cite this article as: Hu *et al.*: MLGO: phylogeny reconstruction and ancestral inference from gene-order data. *BMC Bioinformatics* 2014 **15**:354.