Zhou *et al. BMC Genomics* 2018, **18**(Suppl 7):760
DOI 10.1186/s12864-017-4129-0

BMC Genomics

RESEARCH Open Access

Phylogeny analysis from gene-order data with massive duplications

Lingxi Zhou[4], Yu Lin[2], Bing Feng[4], Jieyi Zhao[3] and Jijun Tang[1,4*]

From 12th International Symposium on Bioinformatics Research and Applications (ISBRA 2016)
Minsk, Belarus. 5-8 June 2016

Abstract

Background: Gene order changes, under rearrangements, insertions, deletions and duplications, have been used as a new type of data source for phylogenetic reconstruction. Because these changes are rare compared to sequence mutations, they allow the inference of phylogeny further back in evolutionary time. There exist many computational methods for the reconstruction of gene-order phylogenies, including widely used maximum parsimonious methods and maximum likelihood methods. However, both methods face challenges in handling large genomes with many duplicated genes, especially in the presence of whole genome duplication.

Methods: In this paper, we present three simple yet powerful methods based on maximum-likelihood (ML) approaches that encode multiplicities of both gene adjacency and gene content information for phylogenetic reconstruction.

Results: Extensive experiments on simulated data sets show that our new method achieves the most accurate phylogenies compared to existing approaches. We also evaluate our method on real whole-genome data from eleven mammals. The package is publicly accessible at http://www.geneorder.org.

Conclusions: Our new encoding schemes successfully incorporate the multiplicity information of gene adjacencies and gene content into an ML framework, and show promising results in reconstruct phylogenies for whole-genome data in the presence of massive duplications.

Keywords: Phylogeny reconstruction, Maximum likelihood, Variable length binary encoding, Whole genome duplication

Background

Phylogeny analysis is one of the key research areas in evolutionary biology. Currently, the dominant data source used in phylogenetic reconstruction is sequence data [1], which can be collected in large amount at low cost (e.g., for coding genes). However, using sequence data (e.g. gene sequences) in phylogenetic reconstruction needs accurate inference of ortholog relationships and provides us only local information – different parts of the genome may evolve according to different evolutionary models, or even be affected by duplications and losses.

Large-scale changes on genomes may hold the key of building a coherent picture of the past history of contemporary species [2]. In such events, entire segments of a genome may be rearranged, duplicated, or deleted. As whole genomes are collected at increasing rates, whole-genome data has become a new and attractive type of data source for phylogenetic analysis [3–8]. Moreover, researchers uncover links between large-scale genomic events (such as rearrangements, duplications, losses leading to copy number variations) and various diseases, especially cancers. Since phylogenetic reconstruction problem is the key to ancestral reconstruction problem, a number of related works [9–14], based on phylogenetic analysis, have been well studied since the 2010s.

*Correspondence: jtang@cse.sc.edu
[1]School of Computer Science and Engineering, Tianjin University, 300072 Tianjin, China
[4]Department of Computer Science and Engineering, University of South Carolina, 29208 Columbia, South Carolina, USA
Full list of author information is available at the end of the article

MPBE [5] and MPME [6] introduced the idea of encoding gene orders into aligned sequences without loss of information. Therefore we can use parsimony software such as TNT [15] and PAUP* [8] developed for molecular sequences to reconstruct gene order phylogeny. Although MPBE and MPME failed to compete with direct parsimonious approaches on whole-genome data [3, 4, 16], they show great speedup and pave the way for future improvements. Moreover, sequence data can be analyzed by searching the phylogeny with maximum likelihood score as suggested by Felsenstein [17] in 1981. Recent algorithmic development and high-performance computing tools such as RAxML [18] have made the maximum likelihood approach feasible for analyzing very large collection of molecular sequences and reconstructing better phylogenies than parsimonious methods. The first successful attempt to use maximum-likelihood to reconstruct a phylogeny from the whole-genome data of bacterials was published [19] in 2011, but that method appeared to be too time-consuming to process eukaryotic genomes. Later, Lin et al. [20] described a maximum-likelihood approach, MLWD, for phylogenetic analysis that takes into account genome rearrangements as well as duplications, insertions, and losses. This MLWD approach can handle high-resolution genomes (with tens of thousands of markers) and can be used in the same analysis for genomes with very different numbers of markers [20]. Although the MLWD method outperforms both distance-based methods [21, 22], the MLWD approach did not make full use of the copy number information of both gene adjacency and gene content, and thus its performance fades out when genomes experienced a large number of duplications, especially in the presence of whole genome duplications.

In this paper, we propose new maximum-likelihood methods for phylogenetic reconstruction from whole-genome data, by taking into account copy number variations in both gene adjacency and gene content. Extensive experiments on simulated data sets showed that our new method achieves the most accurate phylogenies compared to existing approaches. Moreover, we also applied our new method to analyze the real whole-genome data from eleven mammals.

Preliminary

Given a set of n genes labeled as $G = \{1, 2,..., n\}$, we represent a genome by an ordered list of these genes, where each gene may appear more than once in a genome. Given a gene g, we denote its head by g^h and its tail by g^t, with $+g$ indicating that this gene is oriented from tail to head (from g^t to g^h) and $-g$ indicating otherwise (from g^h to g^t). An adjacency of two consecutive genes a and b can form one of the following four possibilities, (a^t, b^h), (a^h, b^h), (a^t, b^t), and (a^h, b^t). A gene c lies at one end of a linear chromosome is called a telomere, denoted by a singleton set (c^t) or (c^h). With the above notations, we can represent a genome by a multiset of adjacencies and telomeres (if there's any). For instance, we represent a simple genome composed of one linear chromosome $(+a, +b, +a, -c, +a)$ and one circular chromosome $(+d, -e)$ as a multiset of adjacencies and telomeres $S = \{(a^t), (a^h, b^t), (b^h, a^t), (a^h, c^h), (c^t, a^t), (a^h), (d^h, e^h), (e^t, d^t)\}$. Note that in the presence of duplicated genes, there is no one-to-one correspondence between genomes and multisets of genes, adjacencies, and telomeres [23]. For example, the genome consisting of the linear chromosome $(+a, -c, +a, +b, +a)$ and the circular one $(+d, -e)$, will have the same multiset of adjacencies and telomeres as the above example.

Genome rearrangements change the ordering of genes on a chromosome and exchange or combine content across chromosomes. An inversion or reversal reverses a segment of genes on a chromosome. A transposition swaps two segments on a chromosome. Translocation breaks at two chromosomes and exchange segments between them. An event of fusion concatenates two chromosomes into one, and a fission event is the reverse and splits one chromosome into two.

Deletion, insertion and duplication not only change the ordering of genes, but also change the copy number of genes. A deletion removes one or a segment of genes from a genome, while insertion adds new genes that have not been present into a chromosome at a time. A segmental duplication copies a single or a segment of genes from a genome, and inserted the copy back to the genome. A whole genome duplication (WGD) accounts for the operation on an ancestral node, by which a genome is transformed into another by duplicating all chromosomes.

Methods

In this section, we first give description of three versions of Variable Length Binary Encoding schemes (VLBE) and then introduce Variable Length Binary Encoding based Phylogeny Reconstruction with Maximum Likelihood on Whole-Genome Data (VLWD*x*).

In the WLMD approach [20], the copy number information of both gene adjacency and gene content has not been fully reflected in the binary encoding. WLMD uses binary encoding to note the absence or presence of an adjacency or gene (i.e., 1 for presence and 0 for absence), but WLMD does not distinguish the number of copies of the same adjacency or gene in the genome.

In this paper, we propose a new encoding scheme that encodes a genome data by Variable Length Binary Encoding schemes (VLBE), which preserves as much as possible of both gene order and gene content information. We then incorporate a dedicated transition model, and develop the phylogenetic reconstruction method, Maximum Likelihood on Whole-Genome Data (VLWDx), which is aimed to be more robust compared to WLMD [20], especially in the presence of a large number of duplications.

For rearrangement-only model, we apply $VLBE_1$ to encode the presence or absence of any adjacency or telomere in the genome. We take into account only the adjacencies and telomeres that appear in at least one of the given genomes. Given n distinct genes in all input genomes is n, there are $\Theta(n^2)$ possible adjacencies and telomeres. However, the number of adjacencies and telomeres that appear in at least one of the input genome is usually much smaller – in fact, it is usually linear in n rather than quadratic [20].

For the general model with not only rearrangements, but also duplications, insertions and deletions, we add the encoding of gene content besides the encoding of adjacencies. For each gene, we apply $VLBE_2$ or $VLBE_3$ to indicate the presence/absence or the multiplicity of this gene in a genome.

In the following three subsections, we give details on the three encoding schemes, along with the resulting encodings for the genome given in Table 1(a).

Variable length binary encoding 1 ($VLBE_1$)

We start with only encoding gene adjacency information. For a dataset D of n genomes, we scan and collect collect all unique adjacencies to obtain a list A of m adjacencies. We count the maximum number of occurrences t for each adjacency $a \in A$ among all the genomes. The encoding of each adjacency a is performed as follows: if genome D_i has k copies of the adjacency a, we append $t - k$ 0's and k 1's to the sequence.

Table 1 (b) gives an example of $VLBE_1$ encoding. We can further reduce the length of these sequences by removing those characters at which every genome has the same state and we do this for the next two encoding schemes.

Variable length binary encoding 2 ($VLBE_2$)

We propose $VLBE_2$ to encode the multiplicity of adjacencies as well as the presence or absence of gene content. For an input dataset D with n genomes, we scan and collect all unique adjacencies to obtain a list A of m adjacencies. We count the maximum number of occurrences t for each adjacency $a \in A$ among all the genomes. We then perform the encoding of each adjacency a as follows: if genome D_i has k copies of the adjacency a, we append $t - k$ 0's and k 1's to the sequence. We also append the encoding of gene content as follows: for each unique gene, if it presents in genome D_i, append 1 at the encoding for genome D_i, otherwise append 0 to the sequence (see Table 2 for an example).

Variable length binary encoding 3 ($VLBE_3$)

We further explore whether variable length binary encoding on gene content would also make a difference on phylogeny reconstruction. $VLBE_3$ is aimed at encoding both adjacencies and gene content. For a dataset D with n genomes, we scan and collect all unique adjacencies to build a list A of m adjacencies. We count the maximum number of occurrences t for each adjacency $a \in A$ and encode each adjacency a as follows: if genome D_i has k copies of adjacency a, we append $t - k$ 0's and k 1's to the encoding sequence for D_i. We also append content encoding in the same way as for the adjacencies. See Table 3 for an example of $VLBE_3$ encoding.

Build phylogeny from sequences

As mentioned above, $VLBE_1$, $VLBE_2$ and $VLBE_3$ aim at transforming gene order information to binary sequences without losing important genomic information, after encoding. The key of phylogenetic reconstruction based on binary encoding is to determine the transition model of flipping a state (from 1 to 0 or from 0 to 1). In order to perform a fair comparison with MLWD, we use the same transition model as described in MLWD [20] here.

Once we build the encoding sequences for all of the input genomes, we use RAxML (version 7.2.8) to reconstruct a tree from these sequences. Although our VLBE encoding may generate a sequence longer than that from other encoding methods mentioned above (up to 2-3 times in all of our experiments), it didn't significantly increase the running time of RAxML, thanks to RAxML's excellent implementation on parallel coding.

Results

Experiments design

We set to evaluate the performance of our approaches on simulated datasets with known "ground truth". We further

Table 1 Example of the binary encoding through $VLBE_1$, for three genomes: G_1: (-2, -1, -3), G_2: (-1, 4, 2), and G_3: (-2, -1, -4, 1, 2)

	Adjacencies							
Encoding	(-3,-2)	(-2,-1)	(-1,-3)	(2,-1)	(-1,4)	(4,2)	(2,-2)	(-1,-4)
G_1	1	01	1	0	0	0	0	0
G_2	0	00	0	1	1	1	0	0
G_3	0	11	0	0	1	0	1	1

Note that (1,2) and (-2,-1) are the same adjacency

Table 2 Example of the binary sequences using *VLBE*$_2$, for three genomes: G_1: (-2, -1, -3), G_2: (-1, 4, 2), and G_3: (-2, -1, -4, 1, 2)

	Adjacencies								Content			
Encoding	(-3,-2)	(-2,-1)	(-1,-3)	(2,-1)	(-1,4)	(4,2)	(2,-2)	(-1,-4)	1	2	3	4
G_1	1	01	1	0	0	0	0	0	1	1	1	0
G_2	0	00	0	1	1	1	0	0	1	1	0	1
G_3	0	11	0	0	1	0	1	1	1	1	0	1

Note that (1,2) and (-2,-1) are the same adjacency

tested our new method on a data set of 11 mammal genomes obtained from Ensembl [24].

We follow the standard practice to set up our simulations [7]. We generate model trees (true trees) with different topologies, then simulate a root genome of n genes and perform random evolutionary events (including rearrangements, duplications, insertions and deletions) along each branch to generate child genomes from the root to obtain datasets of leaf genomes. We then reconstruct trees by applying different methods and compare the results against the known evolutionary history.

The simulation process is carried out as follows. First, we produce a birth-death tree T, which obeys the same way as described in [20]. Then we find the longest path between two leaf nodes, with length = K. We apply different evolutionary rates $r \in \{1, 2, 3, 4\}$ so that the tree diameters are in the range of $d \in \{1n, 2n, 3n, 4n\}$: larger diameter means a genome is more distant from its ancestor, and hence more computationally expensive this data set will be. By timing $1/K$ to tree diameter, we then get the length for a certain branch and we apply a variation coefficient to each branch in this way to vary the length of each branch: for each branch we sample a number s uniformly from the interval $(-1, 1)$ and multiply the branch length by e^s. Thus, a branch would get its length L get by,

$$L = r \times n \times (1/K) \times e^s$$

For evolving on each branch, we use a set of evolutionary events, including inversions, fusions, fissions, translocations, indels, segment duplications and whole genome duplications. During the simulation process, each event is assigned a specific value of probability to be selected.

We compare the accuracy of three different approaches, $VLWD_1$, $VLWD_2$, $VLWD_3$ and MLWD. $VLWD_x$ (Variable Length Encoding Whole Genome Data, which corresponds to the encoding schemes $VLBE_x$) is our new approach; MLWD (Maximum Likelihood on Whole-genome Data) is currently the best available method that scales up to analyze thousands of genes and hundreds of leaves.

Simulation under general model without duplications

We simulate different parameter settings to test our proposed method, and run both our methods and MLWD. Our method outperforms MLWD in every data setting and the improvement is even more significant when the tree diameter gets larger. This result is in line with the observation that variable length binary encoding preserves more adjacency and gene content information than MLWD does.

Figure 1 shows error rates for different methods. The x axis indicates the tree diameter and the y axis indicates the RF error rates, which reflects the percentage of different internal edges between two phylogenetic trees [25].

These simulations show that our VLWD approach can reconstruct more accurate phylogenies from genome data experienced various evolutionary events, than the previous binary encoding-based approach MLWD. $VLWD_3$ also outperforms $VLWD_1$ and $VLWD_2$, indicating the importance of encoding the multiplicity of both adjacencies and gene content.

Simulation under general model with duplications

We generate data sets under a more realistic setting for evolutionary event as well as the genome content. For

Table 3 Example of binary sequences using *VLBE*$_3$, for three genomes: G_1: (-2, -1, -3), G_2: (-1, 4, 2), and G_3: (-2, -1, -4, 1, 2)

	Adjacencies								Content			
Encoding	(-3,-2)	(-2,-1)	(-1,-3)	(2,-1)	(-1,4)	(4,2)	(2,-2)	(-1,-4)	1	2	3	4
G_1	1	01	1	0	0	0	0	0	01	01	1	0
G_2	0	00	0	1	1	1	0	0	01	01	0	1
G_3	0	11	0	0	1	0	1	1	11	11	0	1

Note that (1,2) and (-2,-1) are the same adjacency

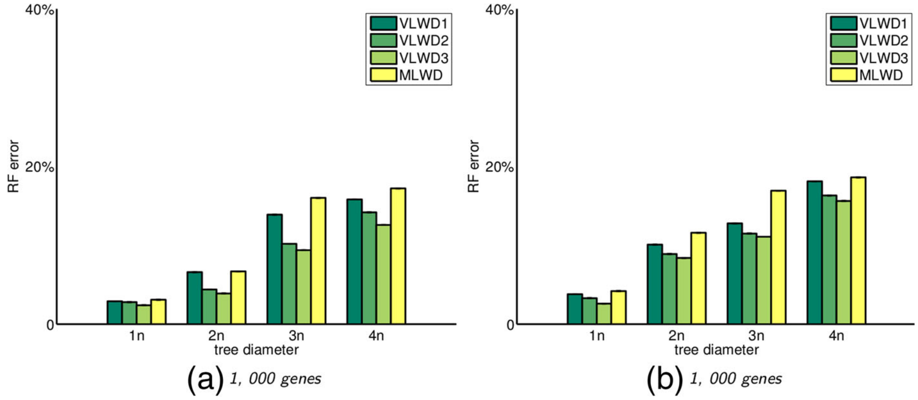

Fig. 1 RF error rates for different approaches for trees with 60 species (left) and 100 species (right), with genomes of 1000 genes and tree diameters from 1 to 4 times the number of genes, under the evolutionary events without duplications VLWD1, VLWD2, VLWD3 are the three proposed methods, and MLWD is the previous encoding method

example, to simulate the evolution of eukaryotic genomes, we generate genome with more than 4,000 genes and the biggest gene family has 20 copies in a single genome.

In our approach, since the encoded sequence of each genome combines information from both gene adjacency and gene content, it is difficult to compute the optimal transition probabilities following the same procedure as described in [20]. Thus we set 1000 as the default bias ratio in the above transition model.

Figures 2 and 3 show the RF error rates. All *VLWD* methods again outperform MLWD, and $VLWD_3$ always maintains the best performance. Figures 2 and 3 together indicate that MLWD returns similar results for data set with and without whole genome duplication, while $VLWD_3$ takes advantage of encoding the multiplicity of both gene adjacencies and gene content, and thus improves on the cases with whole genome duplication compared to those without whole genome duplication.

$VLWD_3$ phylogeny for eleven mammal genomes

In the previous part, we test our $VLWD_3$ approach on simulated data set and achieve very good performance for reconstructing phylogenies. Here we test $VLWD_3$ on the whole genome data of eleven mammal species from online database Ensembl [24].

To obtain the whole genome data of eleven mammal species, we first encode all of the genes into gene orders by using the same gene order to represent all of the homologous genes across different mammal genomes (each genome may contain multiple copies of homologous genes). Subsequently, we input the gene order content and adjacencies into the $VLWD_3$ approach to reconstruct the phylogenetic relationship for these eleven mammal species (see Fig. 4). Thanks to the efficient implementation of RAxML [18], the running time of $VLWD_3$ is similar to *MLWD* [20] and $VLWD_3$ only takes less than ten minutes for the $VLWD_3$ to output the final solution.

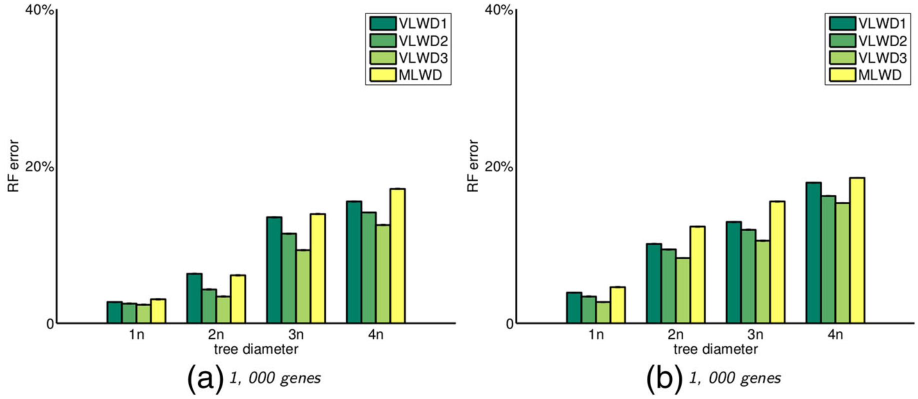

Fig. 2 RF error rates for different approaches for trees with 60 species (left) and 100 species (right), with genomes of 1000 genes and tree diameters from 1 to 4 times the number of genes, under the evolutionary events with segmental duplications VLWD1, VLWD2, VLWD3 are the three proposed methods, and MLWD is the previous encoding method

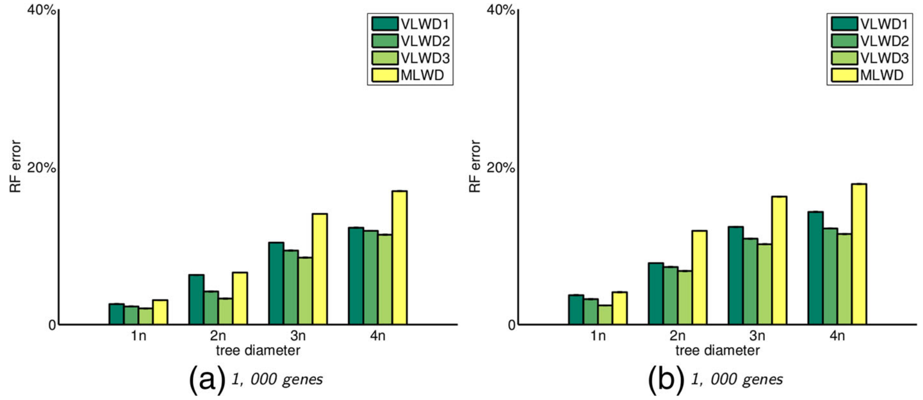

Fig. 3 RF error rates for different approaches for trees with 60 species (left) and 100 species (right), with genomes of 1000 genes and tree diameters from 1 to 4 times the number of genes, under the evolutionary events with both segmental and whole genome duplications VLWD1, VLWD2, VLWD3 are the three proposed methods, and MLWD is the previous encoding method

We compare the $VLWD_3$ phylogeny with the NCBI taxonomy, As Fig. 4 showing, our $VLWD_3$ approach correctly assign the Macaca mulatta and Macaca fascicularis into the Macaca genus and assign the Pan troglodytes and Gorilla gorilla into the Homininae genus. The Rattus norvegicus and Mus musculus are also been correctly assigned into the subfamily Murinae. The Ovis aries and Bos taurus are also been correctly assigned to the Bovidae family. We also compare this $VLWD_3$ phylogeny with the previous gene order based mammal phylogeny study of Luo et al. [26]. There are eight mammal species shared by these two phylogenies, and all of the shared branches for these eight species agree with each other. Moreover, two lowest bootstrap scores (68, 71) on the middle two branches in the tree of Fig. 4 reflect the current controversial opinions in placing primates closer to rodents or carnivores [27–32].

Conclusions

We describe three simple yet powerful approaches for phylogenetic reconstruction based on maximum-likelihood (ML), and design experiments to show the importance of taking into account multiplicities of both gene adjacencies and gene content information. Extensive experiments on simulated data sets show that our proposed approaches achieve the most accurate phylogenies compared to existing methods, particularly in the presence of a large number of duplications or whole genome

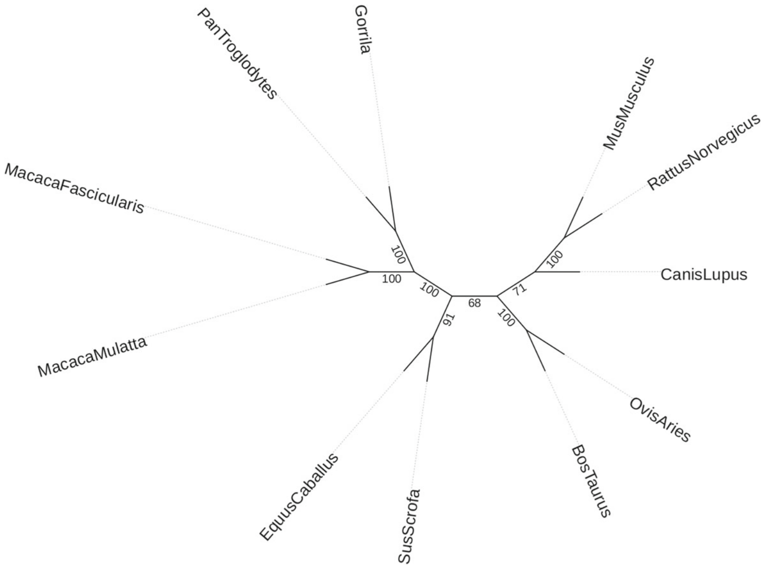

Fig. 4 Phylogeny reconstructed by VLWD for eleven mammal genomes, with bootstrap values shown on branches

duplication. Moreover, we applied our new approach to reconstruct the phylogeny of 11 mammal genomes, using only the whole-genome data from Ensembl [24].

Our new encoding schemes successfully model the multiplicities of gene adjacencies and gene content and incorporate them into a maximum-likelihood framework. Experiments on both simulated and real datasets show the effectiveness and efficiency of our approaches in reconstruction phylogenies using whole-genome data, in the presence of massive duplications.

Acknowledgments

We thank Yan Zhang and Jun Zhou for their support in correcting grammar for this paper. The 2-page abstract has been published in Lecture notes in computer science: Bioinformatics research and applications [33].

Funding

This work (including the source of publication funding) is supported by NSF award #1161586 to LZ, BF and JT.

Availability of data and materials

Our software package is publicly accessible at http://www.geneorder.org.

About this supplement

This article has been published as part of *BMC Genomics* Volume 18 Supplement 7, 2017: Selected articles from the 12th International Symposium on Bioinformatics Research and Applications (ISBRA-16): genomics. The full contents of the supplement are available online at https://bmcgenomics.biomedcentral.com/articles/supplements/volume-18-supplement-7.

Authors' contributions

LZ, YL, JZ and JT work on the algorithm design; LZ and YL work on the experimental design; LZ, YL and BF work on the data preparation and analysis. All authors read and approved the final manuscript.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Author details

[1]School of Computer Science and Engineering, Tianjin University, 300072 Tianjin, China. [2]Research School of Computer Science, Australian National University, 2601 Canberra, ACT, Australia. [3]University of Texas School of Biomedical Informatics at Houston, 77030 Houston, Texas, USA. [4]Department of Computer Science and Engineering, University of South Carolina, 29208 Columbia, South Carolina, USA.

Published: 16 October 2017

References

1. Felsenstein J, Felenstein J. Inferring phylogenies. Sunderland: Sinauer Associates; 2004.
2. Fertin G. Combinatorics of genome rearrangements. Cambridge: MIT press; 2009.
3. Bader D, Moret B, Warnow T, Wyman S, Yan M. GRAPPA (Genome Rearrangements Analysis under Parsimony and other Phylogenetic Algorithms). www.cs.unm.edu/~moret/GRAPPA/.
4. Bourque G, Pevzner PA. Genome-scale evolution: reconstructing gene orders in the ancestral species. Genome Res. 2002;12(1):26–36.
5. Cosner M, Jansen R, Moret B, Raubeson L, Wang L, Warnow T, et al. A new fast heuristic for computing the breakpoint phylogeny and a phylogenetic analysis of a group of highly rearranged chloroplast genomes. In: Proceedings of the 8th International Conference on Intelligent Systems for Molecular Biology. San Diego; 2000. p. 104–15.
6. Moret BM, Wang LS, Warnow T, Wyman SK. New approaches for reconstructing phylogenies from gene order data. Bioinformatics. 2001;17(suppl 1):S165–S173.
7. Edwards A, Nei M, Takezaki N, Sitnikova T, et al. Assessing molecular phylogenies. Science. 1995;267(5195):253.
8. Swofford DL. PAUP 4.0: Phylogenetic analysis using parsimony (and other methods). Sunderland. 1999.
9. Hu F, Lin Y, Tang J. MLGO: phylogeny reconstruction and ancestral inference from gene-order data. BMC Bioinforma. 2014;15(1):1.
10. Zhou L, Hoskins W, Zhao J, Tang J. Ancestral reconstruction under weighted maximum matching. In: Bioinformatics and Biomedicine (BIBM), 2015 IEEE International Conference on. Washington, D.C: IEEE; 2015. p. 1448–55.
11. Hu F, Zhou L, Tang J. In: Cai Z, Eulenstein O, Janies D, Schwartz D, editors. Reconstructing Ancestral Genomic Orders Using Binary Encoding and Probabilistic Models. Springer Berlin Heidelberg: Berlin, Heidelberg; 2013. p. 17–27.
12. Hu F, Zhou J, Zhou L, Tang J. Probabilistic reconstruction of ancestral gene orders with insertions and deletions. Comput Biol Bioinforma, IEEE/ACM Trans. 2014;11(4):667–72.
13. Zhou L, Feng B, Yang N, Tang J. Ancestral reconstruction with duplications using binary encoding and probabilistic model. In: Proceedings of 7th International conference on Bioinformatics and Computational Biology (BICoB). Honolulu; 2015. p. 97–104.
14. Yang N, Hu F, Zhou L, Tang J. Reconstruction of ancestral gene orders using probabilistic and gene encoding approaches. PloS ONE. 2014;9(10): e108796.
15. Goloboff PA, Farris JS, Nixon KC. TNT, a free program for phylogenetic analysis. Cladistics. 2008;24(5):774–86.
16. Xu AW, Moret BME. In: Przytycka TM, Sagot MF, editors. GASTS: Parsimony Scoring under Rearrangements. Springer Berlin Heidelberg: Berlin, Heidelberg; 2011. p. 351–63.
17. Felsenstein J. Evolutionary trees from DNA sequences: a maximum likelihood approach. J Mol Evol. 1981;17(6):368–76.
18. Stamatakis A. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics. 2006;22(21):2688–90.
19. Hu F, Gao N, Zhang M, Tang J. Maximum likelihood phylogenetic reconstruction using gene order encodings. In: 2011 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB). Paris: IEEE; 2011. p. 1–6.
20. Lin Y, Hu F, Tang J, Moret BM. Maximum likelihood phylogenetic reconstruction from high-resolution whole-genome data and a tree of 68 eukaryotes. In: Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing. USA: NIH Public Access; 2013. p. 285.
21. Lin Y, Rajan V, Moret BM. Bootstrapping phylogenies inferred from rearrangement data. Algoritm Mol Biol. 2012;7(1):1.
22. Lin Y, Rajan V, Moret BM. TIBA: a tool for phylogeny inference from rearrangement data with bootstrap analysis. Bioinformatics. 2012;28(24): 3324–5.
23. Lin Y, Moret BM. A new genomic evolutionary model for rearrangements, duplications, and losses that applies across eukaryotes and prokaryotes. J Comput Biol. 2011;18(9):1055–64.
24. Cunningham F, Amode MR, Barrell D, Beal K, Billis K, Brent S, et al. Ensembl 2015. Nucleic Acids Res. 2015;43(D1):D662–D669.
25. Robinson DF, Foulds LR. Comparison of phylogenetic trees. Math Biosci. 1981;53(1-2):131–47.
26. Luo H, Arndt W, Zhang Y, Shi G, Alekseyev MA, Tang J, et al. Phylogenetic analysis of genome rearrangements among five mammalian orders. Mol Phylogenet Evol. 2012;65(3):871–82.
27. Madsen O, Scally M, Douady CJ, Kao DJ, DeBry RW, Adkins R, et al. Parallel adaptive radiations in two major clades of placental mammals. Nature. 2001;409(6820):610–4.

28. Murphy WJ, Eizirik E, Johnson WE, Zhang YP, Ryder OA, O'Brien SJ. Molecular phylogenetics and the origins of placental mammals. Nature. 2001;409(6820):614–8.
29. Amrine-Madsen H, Koepfli KP, Wayne RK, Springer MS. A new phylogenetic marker, apolipoprotein B, provides compelling evidence for eutherian relationships. Mol Phylogenet Evol. 2003;28(2):225–40.
30. Huttley GA, Wakefield MJ, Easteal S. Rates of genome evolution and branching order from whole genome analysis. Mol Biol Evol. 2007;24(8): 1722–30.
31. Wildman DE, Uddin M, Opazo JC, Liu G, Lefort V, Guindon S, et al. Genomics, biogeography, and the diversification of placental mammals. Proc Natl Acad Sci. 2007;104(36):14395–400.
32. Cannarozzi G, Schneider A, Gonnet G. A phylogenomic study of human, dog, and mouse. PLoS Comput Biol. 2007;3(1):e2.
33. Zhou L, Lin Y, Feng B, Zhao J, Tang J. Phylogeny Reconstruction from Whole-Genome Data Using Variable Length Binary Encoding. In: Bioinformatics Research and Applications: 12th International Symposium, ISBRA 2016, Minsk, Belarus, June 5-8, 2016, Proceedings. vol 9683. Berlin Heidelberg: Springer; 2016. p. 345.