MAXIMUM LIKELIHOOD PHYLOGENETIC RECONSTRUCTION FROM HIGH-RESOLUTION WHOLE-GENOME DATA AND A TREE OF 68 EUKARYOTES

YU LIN*

Laboratory for Computational Biology and Bioinformatics, EPFL, Lausanne VD, CH-1015, Switzerland
**E-mail: yu.lin@epfl.ch*

FEI HU and JIJUN TANG

Department of Computer Science and Engineering, University of South Carolina, Columbia, SC 29208, USA
E-mail: {hu5,jtang}@cse.sc.edu

BERNARD M.E. MORET

Laboratory for Computational Biology and Bioinformatics, EPFL, Lausanne VD, CH-1015, Switzerland
E-mail: bernard.moret@epfl.ch

The rapid accumulation of whole-genome data has renewed interest in the study of the evolution of genomic architecture, under such events as rearrangements, duplications, losses. Comparative genomics, evolutionary biology, and cancer research all require tools to elucidate the mechanisms, history, and consequences of those evolutionary events, while phylogenetics could use whole-genome data to enhance its picture of the Tree of Life. Current approaches in the area of phylogenetic analysis are limited to very small collections of closely related genomes using low-resolution data (typically a few hundred syntenic blocks); moreover, these approaches typically do not include duplication and loss events. We describe a maximum likelihood (ML) approach for phylogenetic analysis that takes into account genome rearrangements as well as duplications, insertions, and losses. Our approach can handle high-resolution genomes (with 40,000 or more markers) and can use in the same analysis genomes with very different numbers of markers. Because our approach uses a standard ML reconstruction program (RAxML), it scales up to large trees. We present the results of extensive testing on both simulated and real data showing that our approach returns very accurate results very quickly. In particular, we analyze a dataset of 68 high-resolution eukaryotic genomes, with from 3,000 to 42,000 genes, from the eGOB database; the analysis, including bootstrapping, takes just 3 hours on a desktop system and returns a tree in agreement with all well supported branches, while also suggesting resolutions for some disputed placements.

Keywords: Maximum likelihood; Phylogenetic reconstruction; Genome rearrangement; Gene duplication; Gene loss

1. Introduction

1.1. *Overview*

Phylogenetic analysis is one of the main tools of evolutionary biology. Most of it to date has been carried out using sequence data (or, more rarely, morphological data). Sequence data can be collected in large amounts at very low cost and, at least in the case of coding genes, is relatively well understood, but it requires accurate determination of orthologies and gives us only

local information—and different parts of the genome may evolve at different rates or according to different models. Events that affect the structure of an entire genome may hold the key to building a coherent picture of the past history of contemporary organisms. Such events occur at a much larger scale than sequence mutations—entire blocks of a genome may be permuted (rearrangements), duplicated, or lost. As whole genomes are sequenced at increasing rates, using whole-genome data for phylogenetic analyses is attracting increasing interest, especially as researchers uncover links between large-scale genomic events (rearrangements, duplications leading to increased copy numbers) and various diseases (such as cancer) or health conditions (such as autism). However, using whole-genome data in phylogenetic reconstruction 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.

In this paper, we describe a new approach that resolves these problems and promises to open the way to widespread use of whole-genome data in phylogenetic analysis.

1.2. *Prior work*

Rearrangement data was first used in phylogenetic analysis 80 years ago by Sturtevant and Dobzhansky,[1] but largely ignored for the next 45 years, until revived by Palmer and Thompson[2,3] and Day and Sankoff.[4] In the last 30 years, models of whole-genome evolution, their corresponding distance measures, and algorithms for reconstructing phylogenies under such models, have been the subject of intense research, for which see the text of Fertin *et al.*[5] As in sequence-based phylogenetic reconstruction, approaches based on whole-genome data can be classified in three main categories.

Parsimony-based approaches seek the tree and internal genomes that minimize the total number of events needed to produce the given genomes from a common ancestor. Blanchette *et al.* introduced the first algorithmic approach to the reconstruction of a phylogenetic tree to minimize the total number of *breakpoints*—adjacencies present in one genome, but absent in the other.[6] Moret *et al.* reimplemented this approach in their GRAPPA tool and extended it to *inversion distances*—inversions being the best documented of the hypothesized mechanisms of genomic rearrangements.[7] GRAPPA focused on unichromosomal genomes; to handle multichromosomal genomes, Bourque and Pevzner proposed MGR,[8] based on GRAPPA's distance computations. Whereas BPAnalysis and GRAPPA search all trees and report the one with the best score (an approach that limits GRAPPA to trees of 15 taxa unless combined with the DCM approach of Tang and Moret[9]), MGR uses a heuristic sequential addition method to grow the tree one species at a time. This heuristic approach trades accuracy for scalability, yet MGR does not scale well—in particular, it cannot be used to infer a phylogeny from modern high-resolution data. These various methods are all limited to rearrangements—extensions to handle gene[a] duplications, insertions and losses appear extremely complex and would further limit their scalability.

[a]We use the word "gene" as this is in fact a common form of whole-genome data, but other kinds of markers could be used; more generally, the constituents are syntenic blocks.

Distance-based approaches first estimate the pairwise distances between every pair of leaves, then apply a method such as Neighbor-Joining[10] or FastME[11] to reconstruct the phylogeny from the matrix of pairwise distances. For unichromosomal genomes under inversions, transpositions, and inverted transpositions, Wang and Warnow showed how to estimate a true evolutionary distance from the number of breakpoints.[12,13] For unichromosomal genomes evolving under inversions only, an experimental approach was used by Moret *et al.* to derive an estimate from the inversion edit distance, yielding greatly increased accuracy in tree estimation under both distance and parsimony methods.[14] For multichromosomal genomes, rearrangement operations can be modeled by a single operation called "Double-Cut-and-Join (DCJ)".[15] Lin and Moret developed a procedure to estimate the true evolutionary distance between two genomes under the DCJ model;[16] Lin *et al.* then refined the estimator to include gene duplication and loss events,[17] although that estimator requires knowledge of the direction of time, something usually missing in phylogenetic estimation. The accuracy of distance methods depends entirely on the accuracy of distance estimation and any distance estimator suffers from the saturation problem: as the measured distance increases beyond a certain threshold, the variance in the estimator grows significantly.

Maximum-likelihood (ML) approaches seek the tree and associated model parameters that maximize the probability of producing the given set of leaf genomes. Theoretically, such approaches are much more computationally expensive than both distance-based and parsimony-based approaches, but their accuracy has long been a major attraction in sequence-based phylogenetic analysis. Moreover, in the last few years, packages such as RAxML[18] have largely overcome computational limitations and allowed reconstructions of large trees (with thousands of taxa) and the use of long sequences (to a hundred thousand characters). It was not until last year, however, that the first successful attempt to use ML reconstruction based on whole-genome data was published;[19] results from this study on bacterial genomes were promising, but somewhat difficult to explain, while the method appeared too time-consuming to handle eukaryotic genomes.

2. Methods

Our approach encodes the whole-genome data into binary sequences using both gene adjacencies and gene content, then estimates the transition parameters for the resulting binary sequence data, and finally uses sequence-based ML reconstruction to infer the tree. We call our new approach *Maximum Likelihood on Whole-genome Data (MLWD).*

2.1. *Encoding genomes into binary sequences*

We represent the genome in terms of adjacency information and gene content as follows. Denote the tail of a gene g by g^t and its head by g^h. We write $+g$ to indicate an orientation from tail to head ($g^t \rightarrow g^h$), $-g$ otherwise ($g^h \rightarrow g^t$). Two consecutive genes a and b can be connected by one *adjacency* of one of the following four types: $\{a^t, b^t\}$, $\{a^h, b^t\}$, $\{a^t, b^h\}$, and $\{a^h, b^h\}$. If gene c lies at one end of a linear chromosome, then we have a corresponding singleton set, $\{c^t\}$ or $\{c^h\}$, called a *telomere*. A *genome* can then be represented as a multiset of adjacencies and telomeres. For example, a toy genome composed of one linear chromosome,

	adjacency information						content information			
	$\{a^h, a^h\}$	$\{a^t, b^h\}$	$\{a^t, c^h\}$	$\{b^t, c^t\}$	$\{a^h, d^h\}$	$\{b^t, d^t\}$	a	b	c	d
Genome 1	1	1	1	1	0	0	1	1	1	0
Genome 2	0	1	0	0	1	1	1	1	0	1

$(+a,+b,\text{-}c,+a,+b,\text{-}d,+a)$, and one circular one, $(+e,\text{-}f)$, can be represented by the multiset of adjacencies and telomeres $\{\{a^t\}, \{a^h, b^t\}, \{b^h, c^h\}, \{c^t, a^h\}, \{a^h, b^t\}, \{b^h, d^h\}, \{d^t, a^h\}, \{a^h\}, \{e^h, f^h\}, \{e^t, f^t\}\}$. In the presence of duplicated genes, there is no one-to-one correspondence between genomes and multisets of genes, adjacencies, and telomeres. For example, the genome composed of the linear chromosome $(+a, +b, \text{-}d, +a, +b, \text{-}c, +a)$ and the circular one $(+e, \text{-}f)$, would have the same multisets of adjacencies and telomeres as our toy example.

For data limited to rearrangements (i.e. for genomes with identical gene content), we encode only the adjacency information. For a possible adjacency or telomere, we write 1 (or 0) to indicate its presence (or absence) in a genome. We consider only those adjacencies and telomeres that exist in at least one of the input genomes. If the total number of distinct genes among the input genomes is n, then the total number of distinct adjacencies and telomeres is $\binom{2n+2}{2}$, but the number of adjacencies and telomeres that appear in at least one input genome is typically far smaller—in fact, it is usually linear in n rather than quadratic. For the general model, which includes gene duplications, insertions, and losses in addition to rearrangements, we extend the encoding of adjacencies by also encoding the gene content. For each gene, we write 1 (or 0) to indicate the presence (or absence) of this gene in a genome. For the two toy genomes of Figure 1, the resulting binary sequences and their derivation are shown in Table 1.

2.2. *Estimating transition parameters*

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). Let us first look at adjacencies. Every DCJ operation will select two adjacencies (or telomeres) uniformly at random, and (if adjacencies) break them to create two new adjacencies. Each genome has $n + O(1)$ adjacencies and telomeres ($O(1)$ is the number of linear chromosomes in the genome, viewed as a small constant). Thus the transition probability from 1 to 0 at some fixed index in the sequence is $\frac{2}{n+O(1)}$ under one DCJ operation. 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 general opinion about the probability of eventually breaking an ancestral adjacency (high) vs. that of creating

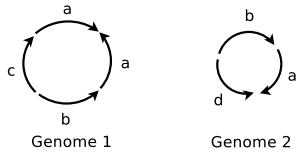


Fig. 1. Two toy genomes.

a particular adjacency along several lineages (low)—a version of homoplasy for adjacencies.

In the general model, we also have transitions for gene content. Once again, the probability of losing a gene independently along several lineages is high, whereas the probability of gaining the same gene independently along several lineages (the standard homoplasy) is low. However, there is no simple uniformity assumption that would enable us to derive a formula for the respective probabilities—there have been attempts to reconstruct phylogenies based on gene content only,[20–22] but they were based on a different approach—so we experimented with various values of the ratio between the probability of a transition from 1 to 0 and that of a transition from 0 to 1.

2.3. *Reconstructing the phylogeny*

Once we have the binary sequences encoding the input genomes and have computed the transition parameters, we use the ML reconstruction program RAxML[18] (version 7.2.8 was used to produce the results given in this paper) to build a tree from these sequences. Because RAxML uses a time-reversible model, it estimates the transition parameters directly from the input sequences by computing the base frequencies. In order to set up the $2n$ ratio, we simply add a direct assignment of the two base frequencies in the code.

3. Results

3.1. *Experimental Design*

We ran a series of experiments on simulated datasets in order to evaluate the performance of our approach against a known "ground truth" under a wide variety of settings. We then ran our reconstruction algorithm on a dataset of 68 eukaryotic genomes, from unicellular parasites to mammalians, obtained from the *Eukaryotic Gene Order Browser (eGOB)* database.[23]

Our simulation studies follow standard practice in phylogenetic reconstruction.[24] We generate model trees under various parameter settings, then use each model tree to evolve an artificial root genome from the root down to the leaves, by performing randomly chosen evolutionary events on the current genome, finally obtaining datasets of leaf genomes for which we know the complete evolutionary history. We then reconstruct trees for each dataset by applying different reconstruction methods and compare the results against the model tree.

3.1.1. *Simulating phylogenetic trees*

A model tree consists of a rooted tree topology and corresponding branch lengths. The trees are generated by a three-step process. We first generate birth-death trees using the tree generator (from the geiger library) in the software R[25] (with a birth rate of 0.001 and a death rate of 0), which simulates the development of a model tree under a uniform, time-homogeneous birth-death process. The branch lengths in such trees are ultrametric (the root-to-leaf paths all have the same length), so, in the second step, the branch lengths are modified as follows. We choose a parameter c; for each branch we sample a number s uniformly from the interval $[-c, +c]$ and multiply the original branch length by e^s (for the experiments in this paper, we set $c = 2$). Thus, each branch length is multiplied by a possibly different random number.

Finally, we rescale all branch lengths to achieve a target diameter D (the length of the longest path, defined as the sum of the edge lengths along that path) for the model tree. (Note that the unit of "length" is one expected evolutionary operation.)

Our experiments are conducted by varying three main parameters: the number of taxa , the number of genes, and the target diameter. We used two values for each of the first two parameters: 50 and 100 taxa, and $1,000$ and $5,000$ genes. For the third parameter, the diameter of the tree, we varied it from n to $4n$, where n is the number of genes. For each setting of the parameters, we generated 100 datasets; data presented below are averages over these 100 datasets.

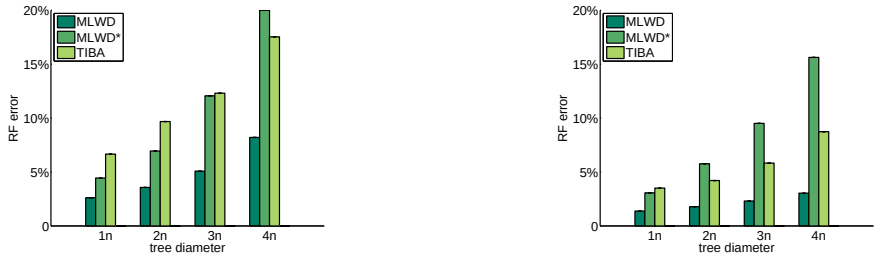
3.1.2. *Simulating evolutionary events along branches in the trees*

In the rearrangement-only model, all evolutionary events along the branches are DCJ operations. The next event is then chosen uniformly at random among all possible DCJ operations.

In the general model, an event can be a DCJ operation or one of a gene duplication, gene insertion, or gene loss. Thus we randomly sample three parameters for each branch: the probability of occurrence of a gene duplication, p_d, the probability of occurrence of a gene insertion, p_i and the probability of occurrence of a gene loss, p_l. (The probability of occurrence of a DCJ operation is then just $p_r = 1 - p_d - p_i - p_l$.) The next evolutionary event is chosen randomly from the four categories according to these parameters. For gene duplication, we uniformly select a position to start duplicating a short segment of chromosomal material and place the new copy to a new position within the genome. We set $L_{\max}$ as the maximum number of genes in the duplicated segment and assume that the number of genes in that segment is a uniform random number between 1 and $L_{\max}$. In our simulations, we used $L_{\max} = 5$. For gene insertion, we tested two different possible scenarios, one for genomes of prokaryotic type and the other for genomes of eukaryotic type. For the former, we uniformly select one position and insert a new gene; for the latter, we uniformly select one existing gene and mutate it into a new gene. Finally, for gene loss, we uniformly select one gene and delete it.

3.2. *Results for simulations under rearrangements*

We compared the accuracy of three different approaches, MLWD, MLWD* and TIBA. MLWD (Maximum Likelihood on Whole-genome Data) is our new approach; MLWD* follows the same procedure as MLWD, but does not use our computation of transition probabilities—instead,



(a) 1,000 genes *(b) 5,000 genes*

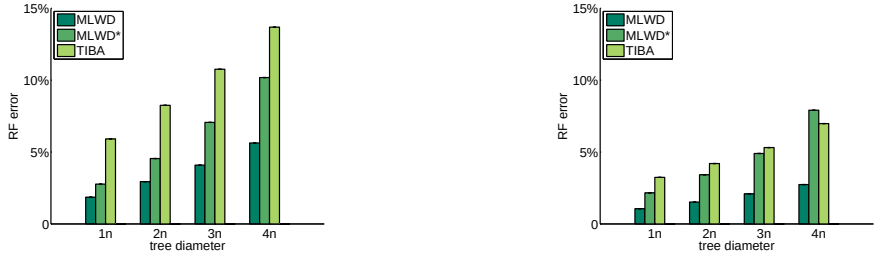
Fig. 2. RF error rates for different approaches for trees with 50 species, with genomes of $1,000$ and $5,000$ genes and tree diameters from one to four times the number of genes, under the rearrangement model.

it allows RAxML to estimate and set them; finally, TIBA is a fast distance-based tool to reconstruct phylogenies from rearrangement data,[26] which combines a pairwise distance estimator[16] and the FastME[11] distance-based reconstruction method. We did not compare with the approaches of Hu *et al.*[19] or those of Cosner *et al.*,[27] because both are too slow and because the former is also limited by their character encodings to a maximum of 20 taxa. Figures 2 and 3 show error rates for different approaches; the x axis indicates the error rates and the y axis indicates the tree diameter. Error rates are RF error rates,[28] the standard measure of error for phylogenetic trees—the RF rate expresses the percentage of edges in error, either because they are missing or because they are wrong.

These representative simulations show that our MLWD approach can reconstruct much more accurate phylogenies from rearrangement data than the distance-based approach TIBA, in line with experience in sequence-based reconstruction. MLWD also outperforms MLWD*, underlining the importance of estimating and setting the transition parameters before applying the sequence-based ML method.

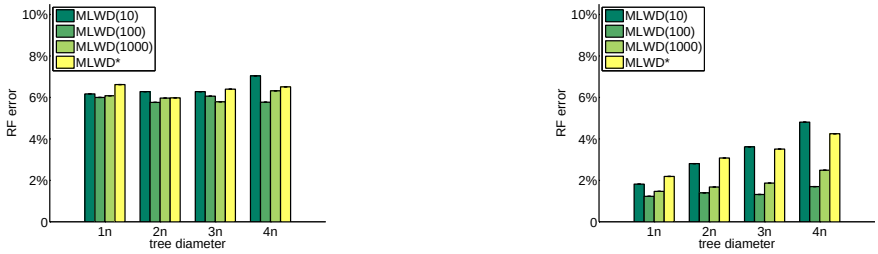
3.3. *Results for simulations under the general model*

Here we generated more complex datasets than for the previous set of experiments. For example, among our simulated eukaryotic genomes, the largest genome has more than 20,000 genes, and the biggest gene family in a single genome has 42 members. In our approach, the encoded sequence of each genome combines both the adjacency and gene content information, which makes it difficult to compute optimal transition probabilities, as discussed in Section 2.2. Thus we set different bias values and compare them under simulation results. If the transition probability of any gene or adjacency from 0 to 1 in MLWD is set to be m times less than that in the opposite direction, we name it MLWD(m) ($m = 10, 100, 1000$). Figure 4 summarizes the RF error rates. Whereas the best ratio in the rearrangement model was $2n$ (as derived in Section 2.2), the best ratio under the general model is much smaller. This difference can be attributed to the relatively modest change in gene content compared to the change in adjacencies: since we encode presence or absence of a gene, but not the number of copies of the gene, not only rearrangements, but also many duplication and loss events will not alter the encoded gene content.



(a) 1,000 genes *(b) 5,000 genes*

Fig. 3. RF error rates for different approaches for trees with 100 species, with genomes of $1,000$ and $5,000$ genes and tree diameters from one to four times the number of genes, under the rearrangement model.



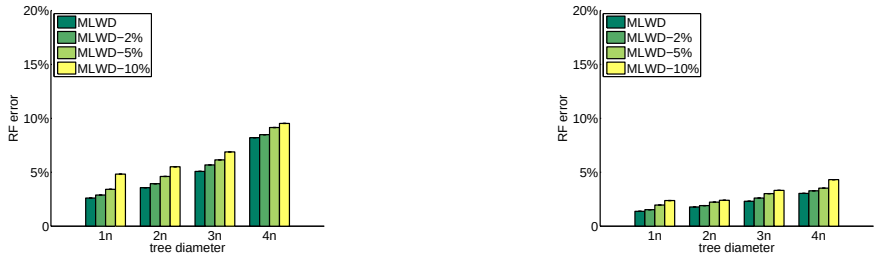
(a) 1,000 genes *(b) 5,000 genes*

Fig. 4. RF error rates for different approaches for trees with 50 species, with initial genomes of size $1,000$ and $5,000$ and tree diameters from one to four times the number of genes in the initial genome, under the general model of evolution.

3.4. *Results for simulated poor assemblies*

High-throughput sequencing has made it possible to sequence many genomes, but the finishing steps—producing a good assembly from the sequence data—are time-consuming and may require much additional laboratory work. Thus many sequenced genomes remain broken into a number of contigs, thereby inducing a loss of adjacencies in the source data. In addition, some assemblies may have errors, thereby producing spurious adjacencies while losing others. We designed experiments to test the robustness of our approach in handling genomes with such assembly defects. We introduce artificial breakages in the leaf genomes by "losing" adjacencies, which correspondingly breaks chromosomes into multiple contigs. For example, MLWD-x% represents the cases of losing x% of adjacencies, that is, x% of the adjacencies are selected uniformly at random and discarded for each genome.

Figure 5 shows RF error rates for MLWD on different quality of genome assemblies under the rearrangement model. Our approach is relatively insensitive to the quality of assembly, especially when the tree diameter is large, that is, when it includes highly diverged taxa. Note that this finding was to be expected in view of the good results of our approach using an encoding that, as observed earlier, does not uniquely identify the ordering of the genes along the chromosomes.



(a) 1,000 genes *(b) 5,000 genes*

Fig. 5. RF error rates for MLWD on different qualities of genome assemblies, for trees with 50 species, with genomes of size $1,000$ and $5,000$. with tree diameters from one to four times the number of genes, under the rearrangement model.

3.5. *Results for a dataset of high-resolution eukaryotic genomes*

Figure 6 shows the reconstructed phylogeny of 68 eukaryotic genomes from the eGOB (Eukaryotic Gene Order Browser) database.[23] The database contains the order information of orthologous genes (identified by OrthoMCL[29]) of 74 different eukaryotic species. The total number of different gene markers in eGOB is around $100,000$. We selected 68 genomes for their size (the number of gene markers) varying from $3k$ to $42k$; the remaining 6 genomes in the database have too few adjacencies (fewer than $3,000$). We encode the adjacency and gene content information of all 68 genomes into 68 binary sequences of length $652,000$. We set the bias ratio to be 100, according to the result of our simulation studies from Section 3.3. Building this phylogeny (using RAxML with fast bootstrapping) took under 3 hours of computing time on a desktop computer.

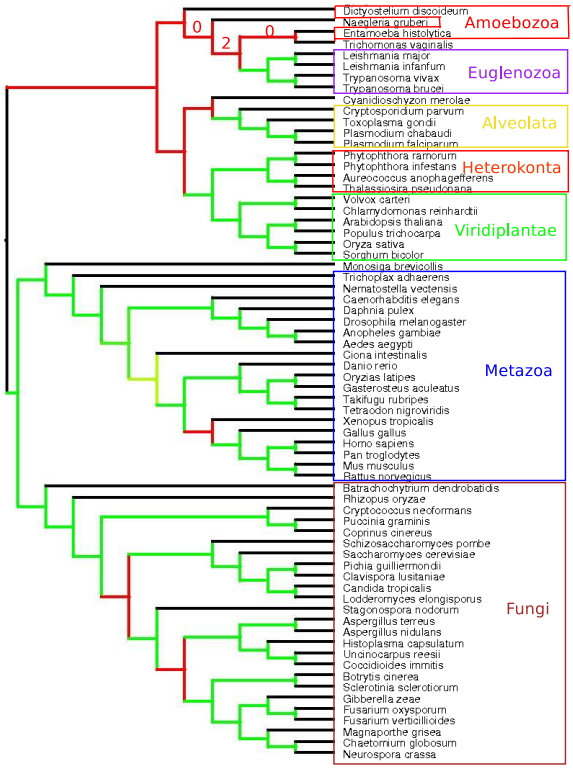


Fig. 6. The reconstructed phylogeny of 68 eukaryotic genomes

The tree is drawn by the tool iTOL;[30] the internal branches are colored into green, yellow and red, indicating, respectively, strong support (bootstrap value > 90), medium support (bootstrap value between 60 and 90), and weak support (bootstrap value < 60). As shown in Figure 6, all major groups in those 68 eukaryotic genomes are correctly identified, with the exception of Amoebozoa. But those incorrect branches with respect to Amoebozoa do receive extremely low bootstrap values (0 and 2), indicating that they are very likely to be wrong. For the phylogeny of Metazoa, the tree is well supported from existing studies.[31,32] For the phylogeny of model fish species (D. rerio, G. aculeatus, O. latipes, T. rubripes, and

T. nigroviridis), two conflicting phylogenies have been published, using different choices of alignment tools and reconstruction methods for sequence data.[33] Our result supports the second phylogeny, which is considered as the correct one by the authors in their discussion.[33] For the phylogeny of Fungi, our results agree with most branches for common species in recent studies.[34,35] It is worth mentioning that among three Chytridiomycota species C. cinereus, P. gramnis, and C. neoformans, our phylogeny shows that C. cinereus and P. gramnis are more closely related, which conflicts with the placement of C. cinereus and C. neoformans as sister taxa, but with very low support value (bootstrapping score 35).[35] C. merolae, a primitive red algae, has been the topic of a longrunning debate over its phylogenetic position.[36] Our result suggests that C. merolae is closer to Alveolata than to Viridiplantae, in agreement with a recent finding obtained by sequencing and comparing expressed sequence tags from different genomes.[37]

Finally, in order to explore the relationship between gene content and gene order, we ran MLWD* on the 68 eukaryotic genomes using only adjacency information as well as using only content information. The tree reconstructed from adjacency information only is poor, with even major clades getting mixed—an unsurprising result in view of the huge variation in gene content among these 68 genomes. The tree reconstructed from gene-content information only correctly identifies all major groups except Amoebozoa; however, it suffers from some major discrepancies with our current understanding of several clades. For example, X. tropicalis is thought to be closer to mammals than to fishes.[38] H. capsulatum, U. reesii, and C. immitis are considered to be in the same order (Onygenales); together with A. nidulans and A. terreus they are considered to be in the same class (Eurotiomycetes), but S. nodorum is thought to belong to a different class (Dothideomycetes).[35] In this particular dataset, which is a sparse sampling of the entire eukaryotic branch of the Tree of Life, most genomes differ significantly in gene content, so that we would expect the tree based on gene-content information to be close to that obtained with both gene adjacencies and gene content. For a denser sampling or for a tree of closely related genomes, adjacency information becomes crucial. A distinguishing feature of MLWD is that it uses both at once and to good effect.

4. Conclusion

In spite of many compelling reasons for using whole-genome data in phylogenetic reconstruction, practice to date has continued to use selected sequences of moderate length using nucleotide-, aminoacid-, or codon-level models. Such models are of course much simpler and much better studied than models for the evolution of genomic architecture. Mostly though, it is the lack of suitable tools that has prevented more widespread use of whole-genome data: previous tools all suffered from serious problems, usually combinations of oversimplified models, poor accuracy, poor scaling, lack of robustness against errors in the data, and lack of any bootstrapping or other statistical assessment procedures.

The approach we presented is the first to overcome all of these difficulties: it uses a fairly general model of genomic evolution (rearrangements plus duplications, insertions, and losses of genomic regions), is very accurate, scales as well as sequence-based approaches, is quite robust against typical assembly errors and omissions of genes, and supports standard bootstrapping

methods. Our analysis of a 68-taxon collection of eukaryotic genomes, ranging from parasitic unicellular organisms with simple genomes to mammals and from around 3,000 genes to over 40,000 genes, could not have been conducted, regardless of computational resources, with any other tools without accepting severe compromises in the data (e.g., equalizing gene content) or the quality of the analysis (by using a distance-based reconstruction method). Our analysis also helps make the case for phylogenetic reconstruction based on whole-genome data. We did not need to choose particular regions of genomes nor to process the data from the eGOB database in any manner; in particular, we did not need to perform a multiple sequence alignment. We were able to run a complete analysis on a "Tree of Life" of all main branches of the Eukaryota, with very divergent genomes (and hence very large pairwise distances), without taking any special precautions and without preinterpreting the data (and thus possibly biasing the output). We could do all of this in a few hours on a desktop machine—in spite of the very long sequences produced by our encoding. We could run the identical software on a collection of organellar genomes or of bacterial genomes with equal success (and in much less time).

Naturally, much work remains to be done. In particular, given the complexity of genomic architecture, current evolutionary models (such as the one we used) are too simple, although even at that level, we need to elucidate simple parameters, such as the ratio of the transition probabilities between loss and gain of a given gene. Using different transition probabilities for adjacencies and for content, by running a compartmentalized analysis, should prove beneficial on large datasets. Larger issues of data preparation also loom. For instance, moving from an assembled genome to the type of data we used continues to require manual intervention—gene-finding, or syntenic block decomposition, are too complex for fully automated procedures. Determination of orthologies, necessary to the identification of syntenic blocks, should be done on the basis of a known phylogeny: that is, the same interdependence exists at the whole-genome level between reconstruction and preprocessing (orthology) as at the sequence level, where it is between reconstruction and alignment. Indeed, most of the methodological questions that the phylogenetic community has been studying in the context of sequence-based reconstruction also arise, in suitably modified terms, in the context of whole-genome data. Our new method provides a first means of empirical enquiry into these questions.

References

1. A. Sturtevant and T. Dobzhansky, *Proc. Nat'l Acad. Sci., USA* **22**, 448 (1936).
2. J. Palmer and W. Thompson, *Proc. Nat'l Acad. Sci., USA* **78**, 5533 (1981).
3. J. Palmer and W. Thompson, *Cell* **29**, 537 (1982).
4. W. Day and D. Sankoff, *J. Theor. Biol.* **127**, 213 (1987).
5. G. Fertin, A. Labarre, I. Rusu, E. Tannier and S. Vialette, *Combinatorics of Genome Rearrangements* (MIT Press, 2009).
6. M. Blanchette, G. Bourque and D. Sankoff, Breakpoint phylogenies, in *Genome Informatics*, eds. S. Miyano and T. Takagi (Univ. Academy Press, Tokyo, 1997) pp. 25–34.
7. B. Moret, S. Wyman, D. Bader, T. Warnow and M. Yan, A new implementation and detailed study of breakpoint analysis, in *Proc. 6th Pacific Symp. on Biocomputing (PSB'01)*, (World Scientific Pub., 2001).
8. G. Bourque and P. Pevzner, *Genome Res.* **12**, 26 (2002).
9. J. Tang and B. 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 Vol. 19 (Oxford U. Press, 2003).
10. N. Saitou and M. Nei, *Mol. Biol. Evol.* **4**, 406 (1987).
11. R. Desper and O. Gascuel, *J. Comput. Biol.* **9**, 687 (2002).
12. L.-S. Wang and T. Warnow, Estimating true evolutionary distances between genomes, in *Proc. 1st Workshop Algs. in Bioinf. (WABI'01)*, Lecture Notes in Comp. Sci.(2149) (Springer Verlag, Berlin, 2001).
13. L.-S. Wang, Exact-IEBP: a new technique for estimating evolutionary distances between whole genomes, in *Proc. 33rd Ann. ACM Symp. Theory of Comput. (STOC'01)*, (ACM Press, New York, 2001).
14. B. Moret, J. Tang, L.-S. Wang and T. Warnow, *J. Comput. Syst. Sci.* **65**, 508 (2002).
15. S. Yancopoulos, O. Attie and R. Friedberg, *Bioinformatics* **21**, 3340 (2005).
16. Y. Lin and B. Moret, Estimating true evolutionary distances under the DCJ model, in *Proc. 16th Int'l Conf. on Intelligent Systems for Mol. Biol. (ISMB'08)*, , Bioinformatics Vol. 24(13)2008.
17. Y. Lin, V. Rajan, K. Swenson and B. Moret, Estimating true evolutionary distances under rearrangements, duplications, and losses, in *Proc. 8th Asia Pacific Bioinf. Conf. (APBC'10)*, , BMC Bioinformatics Vol. 11 (Suppl. 1)2010. S54.
18. A. Stamatakis, *Bioinformatics* **22**, 2688 (2006).
19. F. Hu, N. Gao, M. Zhang and J. Tang, Maximum likelihood phylogenetic reconstruction using gene order encodings, in *Proc. IEEE Symp. Comput. Intell. in Bioinf. & Comput. Biol. (CIBCB'11)*, (IEEE, 2011).
20. B. Snel, P. Bork and M. Huynen, *Nature Genetics* **21**, 108 (1999).
21. D. Huson and M. Steel, *Bioinformatics* **20**, 2044 (2004).
22. H. Zhang, Y. Zhong, B. Hao and X. Gu, *Gene* **441**, 163 (2009).
23. M. López and T. Samuelsson, *Bioinformatics* (2011).
24. D. Hillis and J. Huelsenbeck, *Science* **267**, 255 (1995).
25. R Development Core Team, *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, (2009).
26. Y. Lin, V. Rajan and B. Moret, *J. Comput. Biol.* **18**, 1130 (2011).
27. M. Cosner, R. Jansen, B. Moret, L. Raubeson, L. Wang, T. Warnow and S. Wyman, An empirical comparison of phylogenetic methods on chloroplast gene order data in Campanulaceae, in *Comparative Genomics*, eds. D. Sankoff and J. Nadeau (Kluwer Academic Publishers, Dordrecht, NL, 2000) pp. 99–122.
28. D. Robinson and L. Foulds, *Mathematical Biosciences* **53**, 131 (1981).
29. F. Chen, A. Mackey, J. Vermunt and D. Roos, *PLoS ONE* **2**, p. e383 (2007).
30. I. Letunic and P. Bork, *Nucl. Acids Res.* **39**, W475 (2011).
31. C. Ponting, *Nat. Rev. Genet.* **9**, 689 (2008).
32. M. Srivastava *et al.*, *Nature* **454**, 955 (2008).
33. E. Negrisolo, H. Kuhl, C. Forcato, N. Vitulo, R. Reinhardt, T. Patarnello and L. Bargelloni, *Mol. Biol. Evol.* **27**, 2757 (2010).
34. D. Fitzpatrick, M. Logue, J. Stajich and G. Butler, *BMC Evolutionary Biology* **6**, p. 99 (2006).
35. H. Wang, Z. Xu, L. Gao and B. Hao, *BMC Evolutionary Biology* **9**, p. 195 (2009).
36. H. Nozaki, M. Matsuzaki, M. Takahara, O. Misumi, H. Kuroiwa, M. Hasegawa, T. Shin-i, Y. Kohara, N. Ogasawara and T. Kuroiwa, *J. Mol. Evol.* **56**, 485.
37. F. Burki, K. Shalchian-Tabrizi, M. Minge, A. Skjveland, S. Nikolaev, K. Jakobsen and J. Pawlowski, *PLoS ONE* **2**, p. e790 (2007).
38. U. Hellsten *et al.*, *Science* **328**, 633 (2010).