

# Inferring Genome Rearrangement Phylogeny based on Maximum Likelihood of Gene Adjacencies

Fei Hu<sup>1</sup>, Haiwei Luo<sup>2</sup>, Jian Shi<sup>1</sup>, Yiwei Zhang<sup>1</sup> and Jijun Tang<sup>1\*</sup>

<sup>1</sup>Dept. of Computer Science and Engineering, University of South Carolina  
Columbia, SC 29028, USA

<sup>2</sup>Dept. of Biological Sciences, University of South Carolina  
Columbia, SC 29028, USA

Email: Fei Hu - hufeiy@gmail.com; Haiwei Luo - hluo2006@gmail.com; Jian Shi - shi2@engr.sc.edu; Yiwei Zhang - zhang33@engr.sc.edu; Jijun Tang - jtang@cse.sc.edu;

\*Corresponding author

## Abstract

---

**Background:** Recent technological advances in sequencing and computation have now made it possible to obtain gene content and gene orders within genomes on the order of  $10^9$  base pairs in size or more. These orders can be changed by rearrangement events such as inversions and duplications. Since these events are rare, they enable phylogenetic reconstructions to extend far back in time. Many softwares for the inference of gene order phylogeny have been developed, including widely used maximum parsimony methods such as GRAPPA and MGR. However, these methods have encountered various problems in dealing with today's emerging gene orders. On the other hand, although methods based on maximum likelihood (ML) have been widely used for DNA nucleotide sequence data analysis, the concept of maximum likelihood has not been successfully applied for gene order analysis. In this paper, we provide a simple yet powerful ML method based on binary encoding for phylogenetic reconstruction from gene orders. We call our new method Maximum Likelihood on Binary Encoding (MLBE).

**Results:** We design and conduct a set of experiments to test the performance of our new MLBE method using both simulated and biological datasets. Our results show that MLBE outperforms other methods significantly with larger number of genomes and often return phylogenies without errors. MLBE is also very fast and only needs about several hours to compute datasets with 100 genomes, making it very

suitable for large scale analysis. Our results indicate that MLBE may also be statistically consistent, although thorough investigation should be conducted.

**Conclusions:** Our study of maximum likelihood provides a new direction for genome rearrangement research that avoids the difficulties encountered in the traditional maximum parsimony approaches. Since the encoding deals with only gene adjacencies, it can be easily extended to handle more general cases where gene loss and duplication are common, making the analysis of large nuclear genomes possible.

---

## Introduction

*Genome rearrangements* can be defined as chromosomal structural changes that affect gene order, such as inversions and transpositions. These have been studied intensely since the pioneering papers of Sankoff [1, 2]. Because rearrangements of genes are “rare genomic events” [3], they enable phylogenetic reconstructions to extend far back in time. There are a number of classic papers that apply genome rearrangement phylogeny to relatively simple genomes [4–7]. Genome rearrangement is also at the heart of much work in comparative genomics in organisms with simple and complex genomes [8, 9].

A popular trend in genome rearrangement research is first to develop algorithms that compare two genomes based on given evolutionary models, then extend the results to perform multiple genome comparison. For example, widely used methods such as GRAPPA and MGR are all based on the computation of inversion distance and inversion median, the former compares two genomes and finds the minimum number of inversions between them, while the latter compares three genomes and finds the median genome that minimizes the summation of distances between the median and each of the given ones.

This approach has obvious limitations. Extending pairwise genome comparison to handle more evolutionary events (such as transpositions) is mathematically complicated. Since the median computation under most evolutionary models is very difficult, GRAPPA [10] and MGR [11] are not suitable to handle genomes with hundreds of genes. Even for small genomes, these methods quickly become unusable with the increasing number of events among genomes. GRAPPA and MGR are limited to the number of genomes as well—15 genomes may require several months of computation using GRAPPA. Although recently much improvement has been made on each of the above problems, analyzing dozens of large nuclear genomes is still out of reach.

Thanks to recent algorithmic developments and the introduction of high-performance computation tools such as

RAxML [12], maximum likelihood has become widely used in molecular sequence based phylogenetic reconstruction. However, maximum likelihood approach has not been applied to deal with genome rearrangements, partly because gene orders are generally viewed as as one character with  $2^n n!$  possible states for genomes with  $n$  genes [13]. As a result, it is very difficult to apply the likelihood approach directly on gene orders.

In this paper, we present the first maximum likelihood method for gene order phylogenetic reconstructions. We call this new method Maximum Likelihood on Binary Encoding (MLBE). It is based on encoding that converts gene orders into sequences. We conducted extensive experiments on simulated and biological datasets. Our results show that the new method is very accurate. It may not be the best method when the number of genomes is small and the genomes are closely related. However, for difficult datasets when all the existing methods fail, MLBE still maintains very high accuracy. Our observation also indicates that MLBE may be statistically consistent, although thorough investigation should be conducted.

## Background

### Gene order and genome rearrangements

Given a set of  $n$  genes  $\{g_1, g_2, \dots, g_n\}$ , a genome can be represented by an *ordering* of these genes. To indicate the strandedness of genes, each gene is assigned with an orientation that is either positive, written  $g_i$ , or negative, written  $-g_i$ . Two genes  $i$  and  $j$  are said to be *adjacent* in genome  $G$  if  $i$  is immediately followed by  $j$ , or, equivalently,  $-j$  is immediately followed by  $-i$ . A genome can be *linear* or *circular*. A linear genome is simply a permutation on the multi-subset, while a circular genome can be represented in the same way under the implicit assumption that the permutation closes back on itself.

Let  $G$  be the genome with signed ordering of  $g_1, g_2, \dots, g_n$ . An *inversion* (also called *reversal*) between indices  $i$  and  $j$  ( $i \leq j$ ), produces the genome with linear ordering

$$g_1, g_2, \dots, g_{i-1}, -g_j, -g_{j-1}, \dots, -g_i, g_{j+1}, \dots, g_n.$$

A *transposition* on genome  $G$  acts on three indices  $i, j, k$ , with  $i \leq j$  and  $k \notin [i, j]$ , picking up the interval  $g_i, \dots, g_j$  and inserting it immediately after  $g_k$ . Thus genome  $G$  is replaced by (assume  $k > j$ ):

$$g_1, \dots, g_{i-1}, g_{j+1}, \dots, g_k, g_i, g_{i+1}, \dots, g_j, g_{k+1}, \dots, g_n.$$

An *inverted transposition* is a transposition followed by an inversion of the transposed subsequence; it is also called a *transversion*.

Given two genomes  $\pi_1$  and  $\pi_2$ , we define the *edit distance*  $d(\pi_1, \pi_2)$  as the minimum number of events required to transform one of these genomes into the other. The *inversion distance* between two genomes measures the minimum

number of inversions needed to transform one genome into another. Hannenhalli and Pevzner [14] developed a mathematical and computational framework for signed gene-orders and provided a polynomial-time algorithm to compute inversion distance between two signed gene-orders; Bader et al. [15] later showed that this edit distance can be computed in linear time. Computing the transposition distance is still of unknown computational complexity. The best available method is only a 1.375-approximation [16].

Yancopoulos et al. [17] proposed a universal double-cut-and-join operation that accounts for inversions, translocations, fissions and fusions, which resulted in a new genomic distance that can be computed in linear time. Although there is no direct biological evidence for DCJ operations, these operations are very attractive because they provide a simpler and unifying model for genome rearrangement.

### **Methods for gene order phylogenetic reconstruction**

There exists various methods for reconstructing the gene order phylogenies, including distance-based methods (Neighbor-Joining [18] and FastME [19]), Bayesian method (Badger [20]) and direct optimization methods (GRAPPA [10] and MGR [11]).

Distance-based phylogeny asks for a matrix of pairwise evolutionary distances. Neighbor-Joining (NJ [18]) is one of the most popular distance-based methods which runs in polynomial time and has shown great performance both in synthetic and biological data [21]. For all these distance-based methods that do not have an optimization criteria to follow, the accuracy of the reconstructed tree largely depends on how close the estimated pairwise distance to the true evolutionary distance. That means the estimated pairwise distance is required to be as close to the true distance as possible. *Empirical derived estimation* gives the first shot [22] in which a minimum distance is initially computed between two genomes and an empirical correction is applied based on a statistical model to estimate the true distance. Another more recent approach relies on a structural characterization of a genome pair under double cut and join (DCJ) model [17]. This method (called CDCJ in this study) successfully postpones the emergence of *saturation* [23]. Within the maximum parsimony framework, direct optimization methods such as GRAPPA and MGR explicitly search the best tree with the minimum number of evolutionary events. The first heuristic of such method for gene order phylogeny reconstruction, *BPanalysis*, was proposed by Blanchette *et al.* [24] which seeks to minimize the *breakpoint distance* along all edges, but it is very computationally expensive and limits to analyze small mitochondrial dataset. GRAPPA [10] and MGR [11] make progresses both in running time and accuracy, however since these direct optimization methods run in exponential both in the number of genomes and the number of genes, only relatively small datasets can be handled in an acceptable time.

Several other methods have been proposed. For example, MPBE (Maximum Parsimony on Binary Encoding) [25]

transforms adjacency pairs from the signed permutation into sequence-like strings. These transformed strings are then inputted to the ordinary sequence parsimony software (e.g. PAUP\* 4.0 [26]) to obtain a phylogeny. Afterwards, Wang et al. introduced a new set of encodings (MPME) [27] to improve the accuracy. These encoding based parsimony methods possess slightly better accuracy compared to the NJ method, yet they are computationally very expensive. Consequently, not too much research follows, although they have been shown as a better initialization method in GRAPPA [28], and have been used to estimate ancestral genomes of a set of *Drosophilla* genomes [29]. Our new method MLBE, as the name implies, is inspired by MPBE and uses similar binary encoding. As our experiments shown later, although using similar encoding, using maximum likelihood approach has brought much better accuracy than using maximum parsimonies.

## Methods

### Maximum likelihood based on binary encoding

The maximum likelihood (ML) approach was introduced to reconstruct the phylogenies of DNA nucleotide sequence data by Joe Felsenstein [30]. Maximum likelihood methods have become increasingly popular recently due to the improvement of both speed and accuracy. The key idea behind likelihood based methods is to choose the parameters that maximize the probability of observing the data that we have observed. A crucial advantage of maximum likelihood methods over parsimony methods is the *statistical consistence*: maximum likelihood will converge to the true tree if sufficient long sequences are provided. Felsenstein showed that maximum parsimony is not consistent for molecular sequence data, particularly in the case of unequal evolutionary rates between different lineages [31]. Although the maximum likelihood technique has been used as one of the major methods to infer the phylogenies from DNA nucleotide sequence data, it is yet not feasible to directly apply this technique on gene orders. One possible direction is to decompose the gene orders and convert them into sequences. Various methods have been proposed to make the conversion, among them, the Binary Encoding is the simplest.

Let  $G$  be a signed permutation of  $n$  genes. For linear genomes, genes  $g_0$  and  $g_{n+1}$  are added to indicate the start and end of a genome respectively. For the pair  $(g_i, g_j), 0 \leq i, j \leq n + 1$ , we set up a character to indicate the presence or absence of this adjacency. If  $g_i$  is immediately followed by  $g_i$  in the gene ordering, or  $-g_j$  is immediately followed by  $-g_i$ , we then put a 1 to the sequence at the corresponding site where the character represents this pair and put a 0 otherwise. Although there are up to  $\binom{n+2}{2}$  possible adjacencies, we can further reduce the length of these sequences by removing those characters at which every genome have the same state. These characters can be dropped since either all the genomes have that adjacency or none has it, hence they are not informative for the following phylogeny analysis. The ordering of characters is arbitrary, the same gene pair is represented at the same position for all the

$$\begin{aligned}
G_1 & : (1, 2, 3, 4, 5) \\
G_2 & : (1, 2, -5, -4, 3) \\
G_3 & : (1, -5, -4, -3, -2)
\end{aligned}$$

(a) Three signed linear genomes

	Adjacencies									
	(1, 2)	(2, 3)	(3, 4)	(4, 5)	(5, 6)	(2, -5)	(-4, 3)	(3, 6)	(1, -5)	(-2, 6)
$G_1$	1	1	1	1	1	1	0	0	0	0
$G_2$	1	0	0	0	1	0	1	1	0	0
$G_3$	0	1	1	1	0	0	0	0	1	1

(b) Binary Encoding

Table 1: Examples of the binary encoding.

	Character States									
	$G_1$	A	A	A	G	A	G	G	A	C
$G_2$	A	C	C	T	A	T	T	C	C	T
$G_3$	C	A	A	G	C	T	G	A	A	G

(a) Converted sequences of adjacencies.

Table 2: Examples of the converted sequences.

sequences. Table 1 gives an example of such encoding. Most gene pairs are not shown in this table because they do not appear in any of these genomes. The gene pair 0, 1 occurs in every genome, hence is dropped as well.

The algorithm to determine the encoding can be done by the following two scannings:

- Scan every genome and identify all gene pairs that are presented, store these pairs in a set  $L$ .
- For each gene pair  $p$  in  $L$ , check every genome to determine whether  $p$  is present or not

Given  $N$  genomes with  $n$  genes each, both scans can be implemented easily using arrays. The time complexity is thus  $O(N^2n^2)$ , which works fine for small genomes. However, for large genomes with thousands of genes, the naive implementation is too slow and may take more than a day to process. In our program, we use hash tables to hold all gene pairs in a given genome. As a result, the time to produce encodings is reduced to be within a minute for datasets containing 100 genomes and 10,000 genes.

After converting the gene orders into sequences of 0 and 1, we can compute maximum likelihood phylogenies based on the new sequences. Let  $S$  be the  $N$  sequences transformed from gene order data.  $T$  is the tree with  $N$  leaves where sequence  $S_j$  ( $j = 1, 2, \dots, N$ ) is placed at leaf  $j$ . Let  $L$  be the edge lengths of the tree. Assume we can define and compute a probability  $P(S_1, S_2, \dots, S_n | T, L)$  given a tree and edge lengths, we can describe the maximum likelihood tree inference problem as following: Search over trees of all possible topologies, for each topology  $T$  to

find the lengths  $L$ , that maximize the likelihood  $P(S_1, S_2, \dots, S_n | T, L)$ . The topology and the assignment of edge lengths that give the overall maximum of likelihood score is the desire maximum-likelihood tree.

In this study, instead of developing our own method to search for the maximum likelihood tree, we choose to utilize the power of those widely used packages developed for molecular sequences, such as RAxML [12] and GARLI [32]. These methods require DNA nucleotides (or amino acids) instead of 1 and 0, thus we must convert the labels. One simple way is to code every 1 as nucleotide A, and 0 as nucleotide C. Such coding introduces bias which may impact the inference. As a result, we choose to code every 1 as A or T, and consequently every 0 as C or G. Such choice is a partial random assignment that only asks either A, C pair or T, G pair is used to replace 0, 1 pair during the coding for a certain character so that a balanced number between A, C and T, G are presented in the coding sequences. Table 2 shows the example of nucleotide coding of the genomes presented in Table 1.

We compared several maximum likelihood tools and found that RAxML [12] with General Time Reversible model [33] can efficiently and accurately reconstruct phylogenies from our encoded sequences. RAxML has been steadily improved over the past several years with new techniques such as fast rapid hill climbing algorithm [34] and CAT approximate [35] being introduced. A large-scale performance comparison between RAxML and best competing program (GARLI) proves that RAxML outperforms GARLI with respect to the running time and memory footprint. RAxML also provides a rapid boot strapping search (RBS) [36] that can be combined with a rapid maximum likelihood search and thus allows users to conduct a full maximum likelihood analysis with one single run. The novel RBS is faster than standard bootstrapping search and at the same time yields comparable result. Such full maximum likelihood analysis has been used throughout the phylogeny inference in this study. As we can see in the section of Results, RAxML is very fast and the running time is mostly affected by the number of genomes, instead of the number of genes, hence RAxML can be used without trouble even for large nuclear genomes.

## Simulation setup

We assess topological correctness by computing the *false negatives* (FN) and *false positives* (FP) [37]. Let  $T$  be the true tree and  $T'$  be the tree that we reconstructed. Then the *false negatives* of  $T'$  with respect to  $T$  are those edges in the true tree  $T$  but not in the inferred tree  $T'$ . The *false positives* of  $T'$  with respect to  $T$  are those edges in the inferred tree  $T'$  that do not exist in the true tree  $T$ . In brief the false negatives are the missing edges and false positives are the error edges. These two measures are equal when trees are binary. The *false negatives rate* is the number of false negative divided by the number of internal edges in  $T$ . The *false positives rate* is similarly defined. The *Robinson-Foulds* (RF) rate is defined as the average of the FN and FP rates. An RF rate of more than 5% is generally considered too high [38].

In this paper, we conduct extensive simulations so that the quality of MLBE can be assessed against the known true tree. In our simulations, we generated model tree topologies from the uniform distribution on binary trees, each with 20, 40, 60, 80, 100 leaves. On each tree, we evolved signed permutations of 200 and 1000 genes using various numbers of evolutionary rates: letting  $r$  denote the expected number of inversions along an edge of the true tree, we used values of  $r = 20, 30, \dots, 70$  for 200 genes and  $r = 100, 150, \dots, 350$  for 1000 genes. The actual number of inversions along each edge was sampled from a uniform distribution on the set  $\{\frac{r}{2}, \frac{r}{2} + 1, \dots, \frac{3r}{2}\}$ . To test the robustness of MLBE, we used two evolutionary models: one model allows only inversions, the other uses a mix of 80% inversions and 20% transpositions. For each combination of parameter settings, we ran 20 datasets and averaged the results.

We compared MLBE with other three methods: FastME with the true distance estimator based on DCJ distance (CDCJ) [23], FastME with Empirical Distance Estimation (EDE) [22] and MPBE solved by PAUP\* heuristically [25]. Since our simulated datasets have more than 20 genomes and the genomes are relatively distant, neither GRAPPA nor MGR can finish any of them within days of computation, hence these two methods were excluded.

The problem of most parsimony on binary sequence had been proved to be NP-hard [39]. The size of the dataset in our experiment is way larger than the amount that can be handled in branch and bound search by PAUP\*, therefore we used heuristic search throughout the whole experiment with the auto-increment turned on. As the searches always return a lot of local optima trees, we computed the strict consensus (as implemented in *phylip* [40]) over all the local optima. Then the consensus tree is taken as the “maximum-parsimony tree”. Although the setup of MPBE may be different from that was used by Wang et al. [25], the results are comparable.

### **Biological data preparation**

The whole genomic DNA sequences of 12 *Prochlorococcus* (P. MIT9301, P. MIT9215, P. AS9601, P. MIT9312, P. MIT9515, P. MED4, P. NATL1A, P. NATL2A, P. SS120, P. MIT9211, P. MIT9313, P. MIT9303 and 9 non-diazotrophic *Synechococcus* (S. WH8102, S. CC9605, S. CC9902, S. WH7803, S. CC9311, S. RCC307, S. PCC7942, S. PCC6301, S. PCC7002) strains were downloaded from NCBI and annotated by RAST Server [41, 42]. Using Perl scripts, the predicted protein-coding gene were parsed. Every possible pair of proteomes was assembled with an all-versus-all BLASTP [43] and orthologous genes were predicted using MSOAR software [44]. We obtained 901 shared orthologs among the 21 cyanobacterial genomes, which were used to construct a sequence concatenation tree using Jones-Taylor-Thornton (JTT) model [45] with gamma correction using the PHYLIP [40] and FastME software [Desper2002], and a distance-based gene order tree using the CDCJ distance [23].

## Results

### Simulation results

Figure 1 and Figure 2 show the topological accuracy of four methods: NJ with EDE, NJ with corrected DCJ distance (CDCJ), MPBE and MLBE. Due to format constraints, we only show the results of 20, 60 and 100 genomes.

The performance of MPBE was obviously much worse than the results obtained from MLBE. It is accurate for smaller number of events, but the performance was quickly worsened with increasing number of events and its error rates exceed 30% for most test cases.

The true distance estimator of CDCJ [23] has the best result for smaller number of genomes, but its accuracy quickly decreased with more genomes. On the other hand, the accuracy of MLBE is steady and for almost all test cases, the errors are close to 0%. Our experiment shows that MLBE is the best method for datasets with more than 40 genomes. The trend of these two figures suggests that MLBE is more accurate with larger number of genomes. We observed that the number of informative sites generally increases with the number of genomes. Figure 3 shows the relationship between RF error rates with the number of informative sites. For both 200 genes and 1000 genes, we can see a general trend that the more informative sites, the more accurate MLBE is. Such trend is more obvious for 1000 genes, where many more sites are presented. This figure indicates that MLBE may also be statistical consistent, and further investigation is required to confirm.

The number of genomes, instead of the number of genes has more impact on the computation time of MLBE. A 20 genome dataset required no more than 10 minutes to finish, while a 100 genome dataset required up to eight hours. We also tested some datasets with more than 10,000 genes which can be finished within a day of computation, indicating that MLBE has potential to be used for large nuclear genomes. However we did not conduct systematic testings on these range since for nuclear genomes, gene loss and gene duplication are frequent and MLBE currently cannot handle these events.

### Biological results

We compiled a cyanobacterial dataset consisting of two genus, *Prochlorococcus* (12 genomes) and *Synechococcus* (9 genomes), which was used to study the reductive genome evolution of *Prochlorococcus* (Luo et al. in review). The phylogeny derived from MLBE (Figure 4C) is consistent with the ortholog concatenation tree (Figure 4A) and the CDCJ distance-based gene order tree (Figure 4B). All trees in Figure 4 support that *Prochlorococcus* forms a monophyletic cluster, whereas *Synechococcus* is a paraphyletic group. Of the *Prochlorococcus* genomes, the CDCJ distance-based gene order tree does not resolve the relationship of P. MIT9515 and P. MED4 (Figure 4B). Of the *Synechococcus* genomes, the MLBE tree does not resolve the evolutionary relationship of the clade S. CC9605-S.

CC9902-S. WH8102 and the clade S. CC9311-S. WH7803 (Figure 4C). This result is consistent with those obtained from simulations where neighbor-joining with CDCJ distance has slightly better accuracy for small number of genomes.

## Conclusions

We developed the first maximum likelihood method (MLBE) for gene order phylogenetic reconstructions. Our tests on simulated and biological datasets show that MLBE is very accurate and scales well to accommodate emerging large scale gene orders. Our approach provides a new direction for genome rearrangement research. Since the encoding is based on the presence and absence of gene adjacencies, it will be relatively easy to extend our work so that other evolutionary events such as gene loss and gene duplication can be handled. There exist other more sophisticated encodings and we are also investigating whether these encodings can further improve the accuracy. Our preliminary research on both direction suggests that likelihood methods based on encodings have great potential and formal mathematical and statistical analysis of these methods are desired.

## Authors contributions

All authors contributed to the development and implementation of the method. FH and JT developed the method and conducted the experiment. HL and YZ analyzed the results from real data and simulated data respectively. JS were in charge of designing the simulator.

## Acknowledgments

The authors were supported by US National Institutes of Health (grant number 5R01GM078991-03 and 3R01GM078991-03S1). All experiments were conducted on a 128-core shared memory computer supported by US National Science Foundation grant (NSF grant number CNS 0708391).

## Competing interest

None to declare.

## References

1. Blanchette M, Bourque G, Sankoff D: **Breakpoint Phylogenies**. In *Genome Informatics*. Edited by Miyano S, Takagi T, Tokyo, Japan: University Academy Press 1997:25–34.
2. Sankoff D, Blanchette M: **Multiple genome rearrangement and breakpoint phylogeny**. *Journal of Computational Biology* 1998, **5**:555–570.
3. Rokas A, Holland P: **Rare genomic changes as a tool for phylogenetics**. *Trends in Ecology and Evolution* 2000, **15**:454–459.

4. Downie S, Palmer J: **Use of chloroplast DNA rearrangements in reconstructing plant phylogeny.** In *Plant Molecular Systematics*. Edited by Soltis P, Soltis D, Doyle J 1992:14–35.
5. Raubeson L, Jansen R: **Chloroplast DNA evidence on the ancient evolutionary split in vascular land plants.** *Science* 1992, **255**:1697–1699.
6. Luo H, Shi J, Arndt W, Tang J, Friedman R: **Gene Order Phylogeny of the Genus Prochlorococcus.** *PLoS One* 2008, **3**:e3837.
7. Luo H, Sun Z, Arndt W, Shi J, Friedman R, Tang J: **Gene Order Phylogeny and the Evolution of Methanogens.** *PLoS One* 2009, **4**:e6069.
8. Pevzner P, Tesler G: **Human and mouse genomic sequences reveal extensive breakpoint reuse in mammalian evolution.** *Proceedings of the National Academy of Sciences USA* 2003, **100**:7672–7677.
9. Richards S: **Comparative genome sequencing of *Drosophila pseudoobscura*: Chromosomal, gene and cis-element evolution.** *Genome Research* 2005, **15**:1–18.
10. Moret B, Wyman S, Bader D, Warnow T, Yan M: **A new implementation and detailed study of breakpoint analysis.** In *Proceedings of the 6th Pacific Symp. on Biocomputing (PSB'01)* 2001:583–594.
11. Bourque G, Pevzner P: **Genome-scale evolution: reconstructing gene orders in the ancestral species.** *Genome Research* 2002, **12**:26–36.
12. Stamatakis A, Hoover P, Rougemont J: **A rapid bootstrap algorithm for the RAxML web-servers.** *Syst. Biol.* 2008, **75**:758–771.
13. Moret B, , Warnow T: **Advances in phylogeny reconstruction from gene order and content data.** *Methods in Enzymology* 2005, **395**:673–700.
14. Hannenhalli S, Pevzner P: **Transforming cabbage into turnip (polynomial algorithm for sorting signed permutations by reversals).** In *Proceedings of the 27th Ann. Symp. Theory of Computing (STOC'95)* 1995:99–124.
15. Bader D, Moret B, Yan M: **A Linear-Time Algorithm for Computing Inversion Distance between Signed Permutations with an Experimental Study.** In *Proc. 7th Int'l Workshop on Algorithms and Data Structures (WADS 2001)*, Volume 2125 of Lecture Notes in Computer Science, Providence, RI: Springer-Verlag 2001:365–376.
16. Elias I, Hartman T: **A 1.375-approximation algorithm for sorting by transpositions.** In *Proc 5th Workshop Algs in Bioinformatics, Volume 3692* 2005:204–215.
17. Yancopoulos S, Attie O, Friedberg R: **Efficient sorting of genomic per-mutations by translocation, inversion and block interchange.** *Bioinformatics* 2005, **21**:3340–3346.
18. Saitou N, Nei M: **The neighbor-joining method: A new method for reconstructing phylogenetic trees.** *Mol. Biol. Evol.* 1987, **4**:406–425.
19. Desper R, Gascuel O: **Fast and accurate phylogeny reconstruction algorithms based on the minimum evolution principle.** *J. Comput. Biol.* 2002, **9**:687–705.
20. Larget B, Kadane J, Simon D: **A Bayesian approach to the estimation of ancestral genome arrangements.** *Mol. Phy. Evol.* 2005, **36**:214–223.
21. Wang L, Jansen R, Moret B, Raubeson L, Warnow T: **Distance-based genome rearrangement phylogeny.** *J. Mol. Evol.* 2006, **63**:473–483.
22. Moret B, Wang L, Warnow T, Wyman S: **New approaches for reconstructing phylogenies from gene order data.** *Bioinformatics* 2001, **17**:S165–S173.
23. Lin Y, Moret B: **Estimating true evolutionary distances under the DCJ model.** *Bioinformatics* 2008, **24**:i114–i122.
24. Blanchette M, Bourque G, Sankoff D: **Breakpoint phylogenies.** *Genome Informatics* 1997, :25–34.
25. Cosner M, Jansen R, Moret B, Raubeson L, Wang L, Warnow T, Wyman S: **A new fast heuristic for computing the breakpoint phylogeny and experimental phylogenetic analyses of real and synthetic data.** In *Proceedings of the 8th Intl. Conf. on Intel. Sys. for Mol. Bio. (ISMB'00)* 2000.
26. Swofford D: **PAUP\*. Phylogenetic Analysis Using Parsimony (\*and Other Methods). Version 4.** Sunderland, MA 2003.
27. Moret B, Wang L, Warnow T, Wyman S: **New approaches for reconstructing phylogenies based on gene order.** In *Proceedings of the 9th Intl. Conf. on Intel. Sys. for Mol. Bio. (ISMB'01)* 2001:165–173.

28. Tang J, Wang L: **Improving Genome Rearrangement Phylogeny Using Sequence-Style Parsimony**. *Bioinformatic and Bioengineering, IEEE International Symposium on* 2005, :137–144.
29. Bhutkar A, Russo S, Smith T, Gelbart W: **Chromosomal rearrangement inferred from comparisons of twelve Drosophila genomes**. *Genome Research* 2007, **10.1101/gr.7062307**.
30. Felsenstein J: **Evolutionary trees from DNA sequences: a maximum-likelihood approach**. *Journal of molecular evolution* 1981, **17**:368–376.
31. Felsenstein J: **Cases in which parsimony or compatibility methods will be positively misleading**. *Systematic Zoology* 1978, **27**:401–410.
32. Zwickl D: **Genetic algorithm approaches for the phylogenetic analysis of large biological sequence datasets under the maximum likelihood criterion**. *PhD thesis*, TX University of Texas at Austin 2006.
33. Tavar S: **Some Probabilistic and Statistical Problems in the Analysis of DNA Sequences**. In *Lectures on Mathematics in the Life Sciences, Volume 17* 1986.
34. Stamatakis A, Blagojevic F, Nikolopoulos D, Antonopoulos C: **Exploring New Search Algorithms and Hardware for Phylogenetics: RAxML Meets the IBM Cell**. *The Journal of VLSI Signal Processing* 2007, **48**:271–286.
35. Stamatakis A: **Phylogenetic models of rate heterogeneity: a high performance computing perspective**. In *in Proceedings of IPDPS2006* 2006.
36. Stamatakis A: **RAxML-VI-HPC:Maximum Likelihood-based Phylogenetic Analyses with Thousands of Taxa and Mixed Models**. *Bioinformatics* 2006, **22**:2688–2690.
37. Robinson D, Foulds L: **Comparison of phylogenetic trees**. *Mathematical Biosciences* 1981, **53**:131–147.
38. Swofford D, Olson G, Waddell P, Hillis D: **Phylogenetic inferences**. In *Molecular Systematics*, 2nd edition. Edited by Hillis DM, Moritz C, Mable B 1996.
39. Foulds L, Graham R: **The Steiner problem in phylogeny is NP-complete**. *Advances in Applied Mathematics* 1982, **3**:43–49.
40. Felsenstein J: **PHYLIP-Phylogeny Inference Package**. *Cladistics* 1989, **5**:164–166.
41. Aziz R, Bartels D, Best A, DeJongh M, Disz T, Edwards R, Formsma K, Gerdes S, Glass E, Kubal M, Meyer F, Olsen G, Olson R, Osterman A, Overbeek R, McNeil L, Paarmann D, Paczian T, Parrello B, Pusch G, Reich C, Stevens R, Vassieva O, Vonstein V, Wilke A, Zagnitko O: **The RAST Server: Rapid Annotations using Subsystems Technology**. *BMC Genomics* 2008, **9**:75.
42. Luo H, Shi J, Arndt W, Tang J, Friedman R: **Gene order phylogeny of the genus Prochlorococcus**. *PLoS ONE* 2008, **3**:e3837.
43. Altschul S, Madden T, Schaffer A, Zhang J, Zhang Z, Miller W, Lipman D: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs**. *Nucleic Acids Research* 1997, **25**:3389–3402.
44. Chen X, Zheng J, Fu Z, Nan P, Zhong Y, Lonardi S, Jiang T: **Assignment of Orthologous Genes via Genome Rearrangement**. *IEEE/ACM Trans. Comput. Biol. Bioinformatics* 2005, **2**(4):302–315.
45. Jones D, Taylor W, Thornton J: **The rapid generation of mutation data matrices from protein sequences**. *Comput. Appl. Biosci* 1992, **8**:275–282.

## Figures

### Figure 1 - RF rates for 200 genes

The RF rates on the phylogenies constructed from dataset simulated by different settings.  $r$  is the expected number of events per edge. (left) inversion only model; (right) 80% inversion and 20% transposition.

### Figure 2 - RF rates for 1000 genes

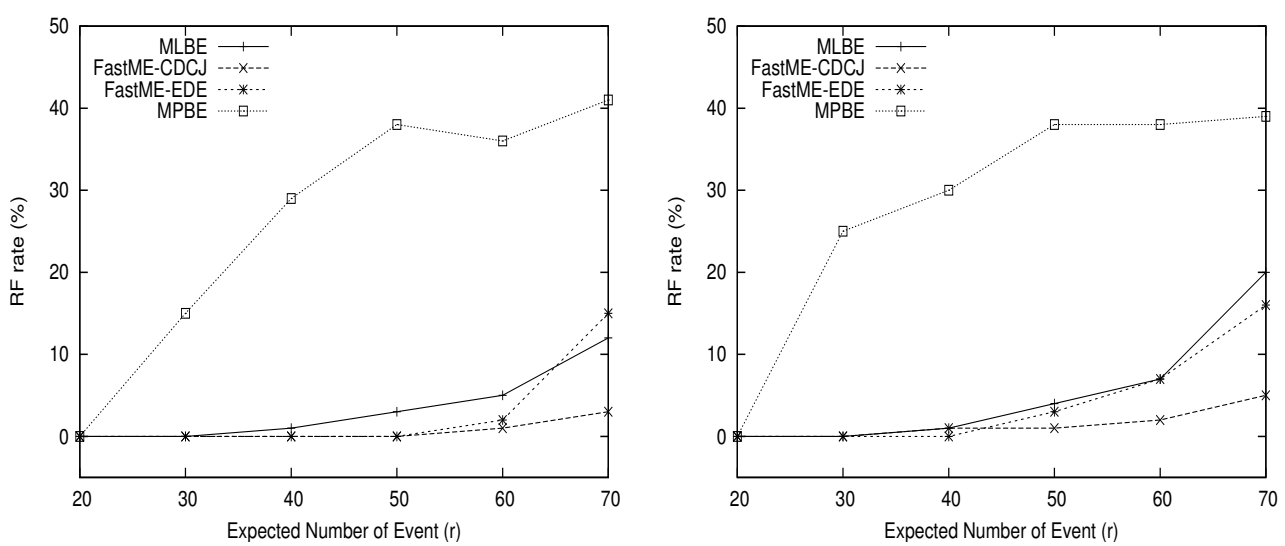
The RF rates on the phylogenies constructed from dataset simulated by different settings.  $r$  is the expected number of events per edge. (left) inversion only model; (right) 80% inversion and 20% transposition.

**Figure 3 - RF rates with respect to the number of informative sites**

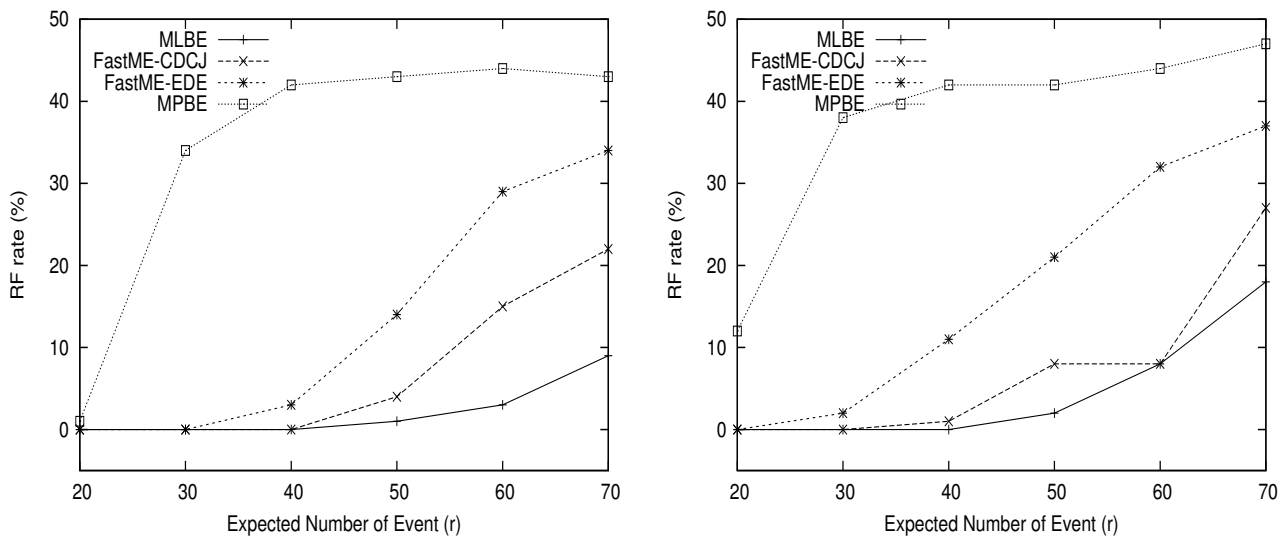
(left) 200 genes; (right) 1000 genes.

**Figure 4 - Phylogenetic relationship of 12 Prochlorococcus and 9 Synechococcus genomes**

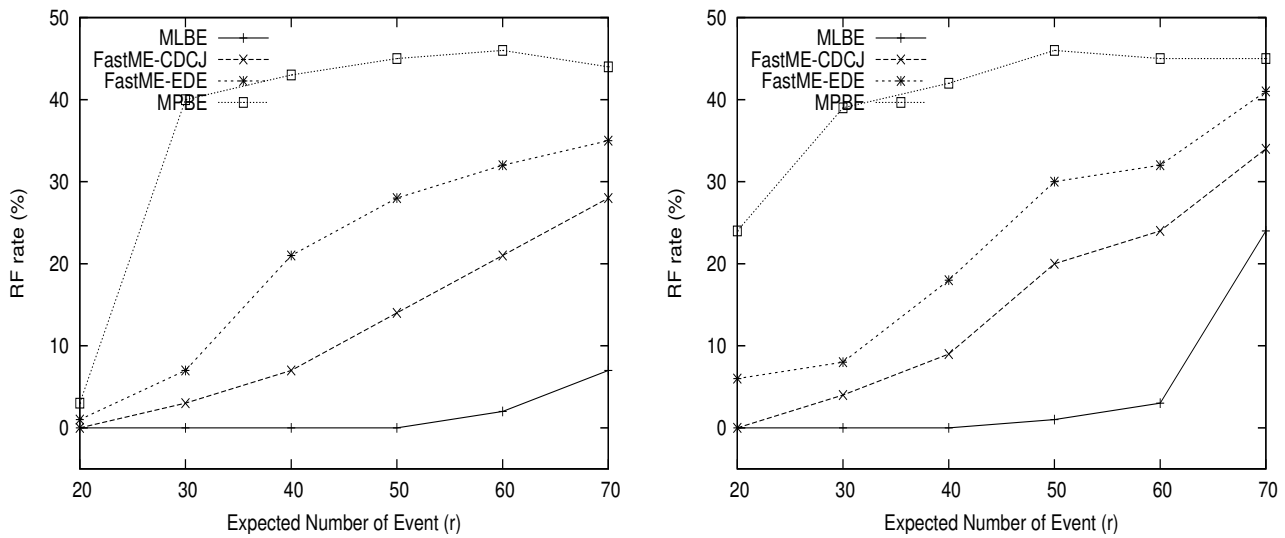
(A) FastME reconstruction of random concatenation of 100 protein sequences sampled from shared orthologs. Values at nodes show the number of times that the clade defined by that node appears in the 100 random concatenation trees; (B) Distance-based gene order phylogeny using CDCJ distance; (C) Tree obtained using MLBE.



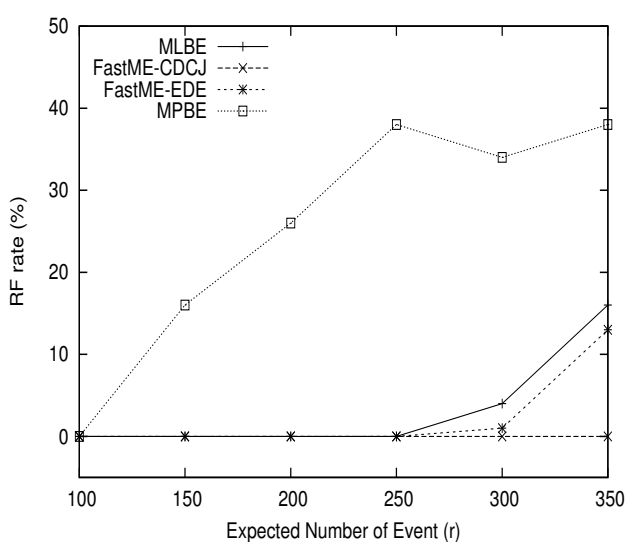
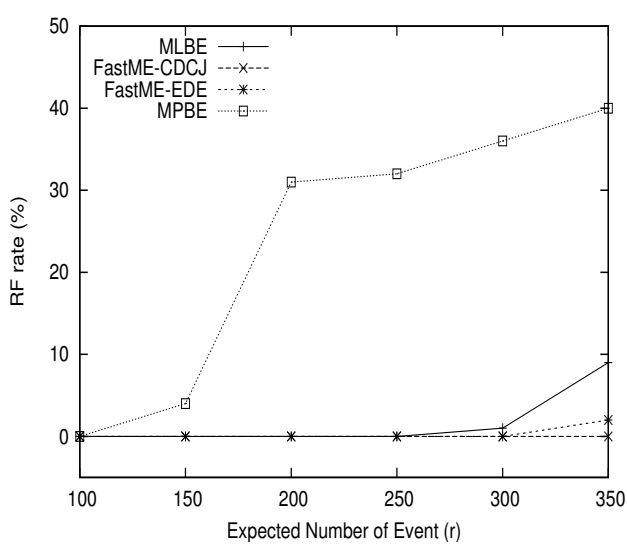
(a) 200 genes, 20 genomes



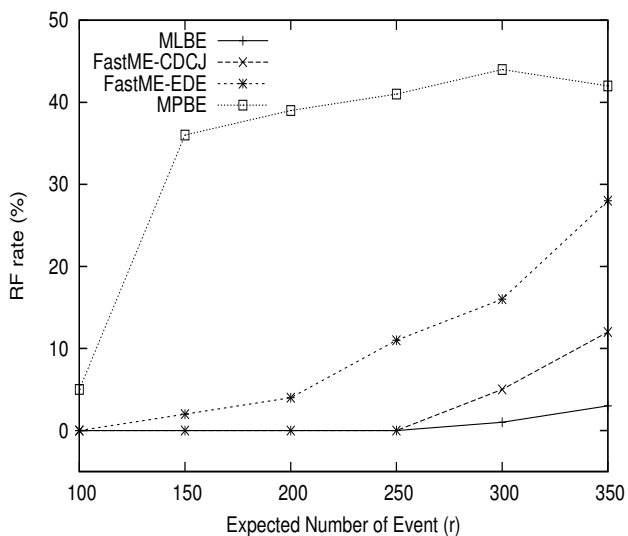
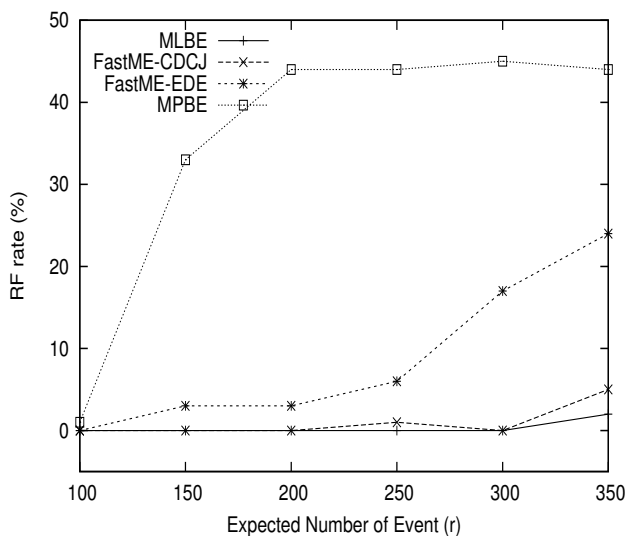
(b) 200 genes, 60 genomes



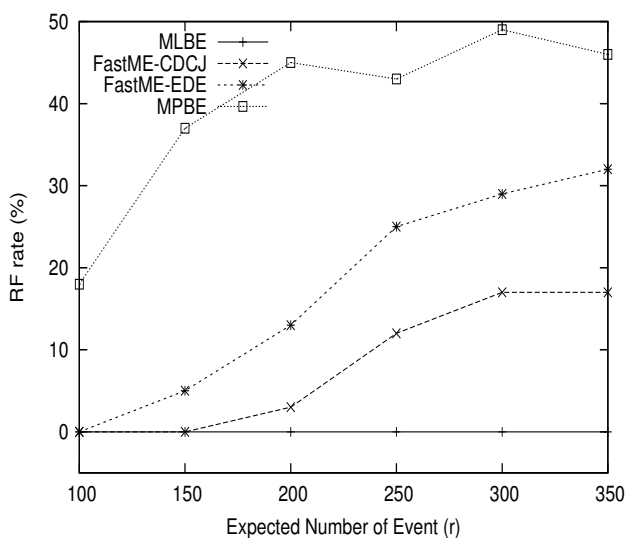
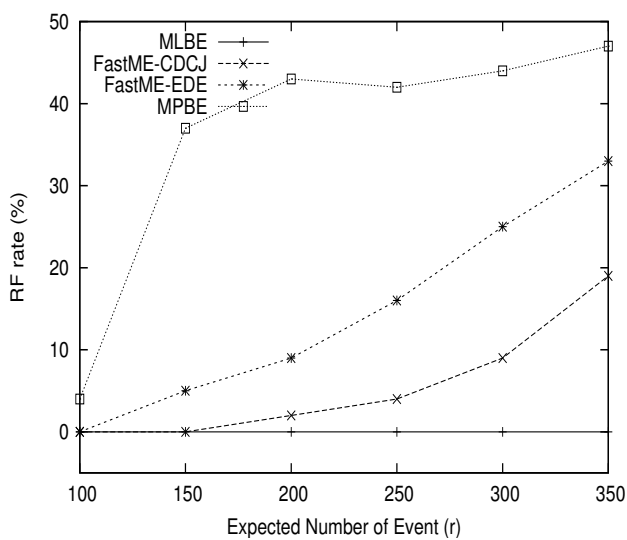
(c) 200 genes, 100 genomes



(a) 1000 genes, 20 genomes



(b) 1000 genes, 60 genomes



(c) 1000 genes, 100 genomes

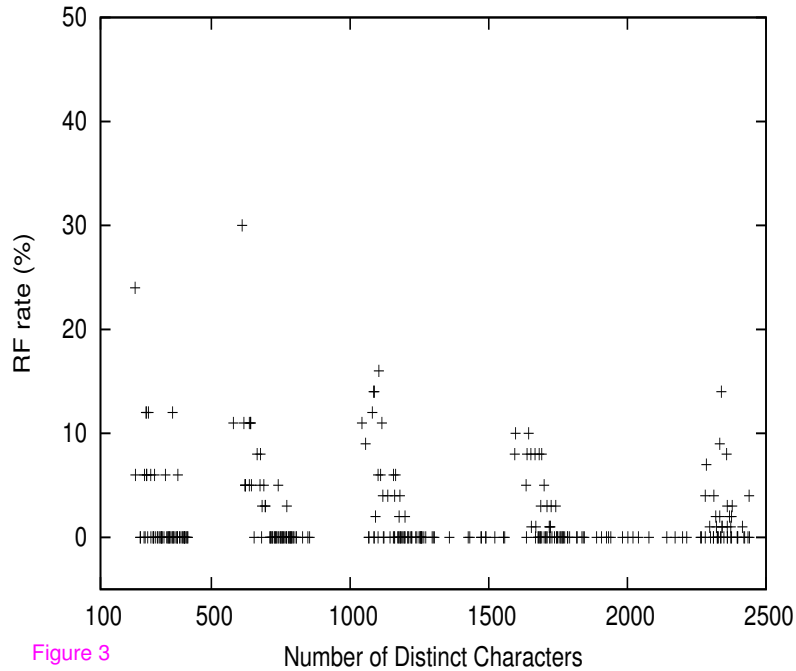
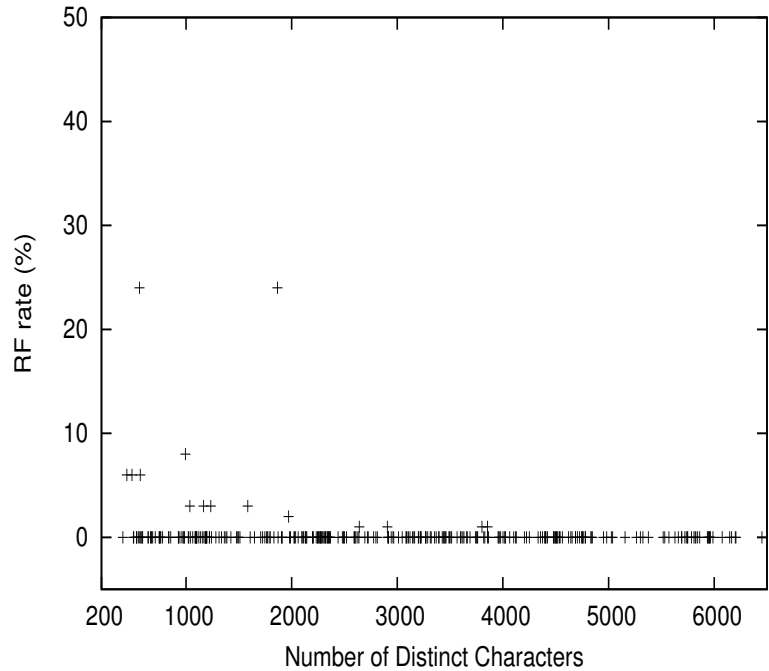


Figure 3



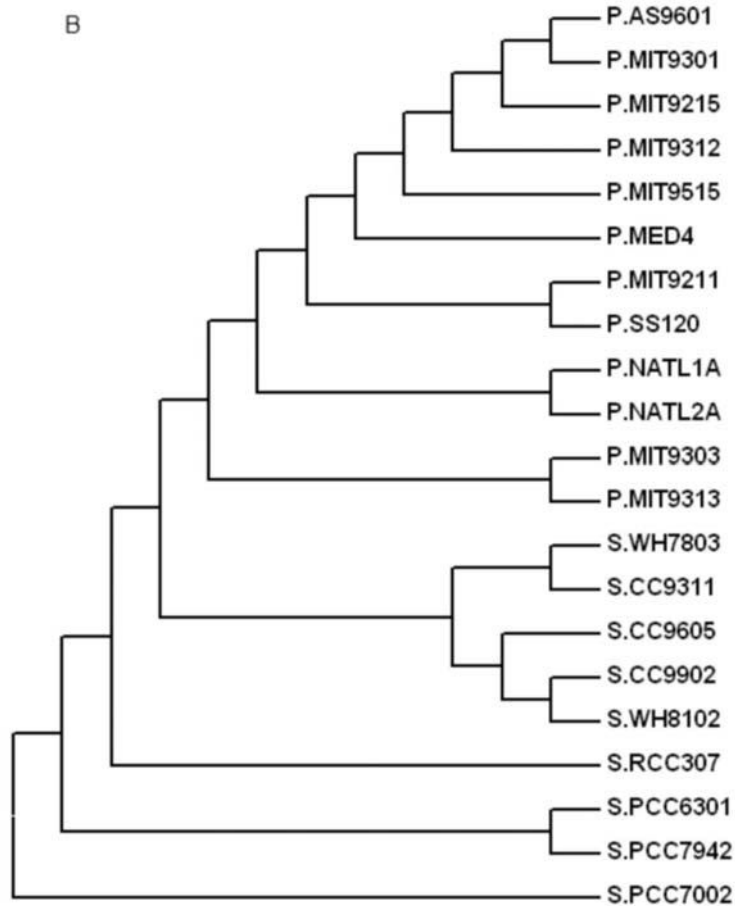
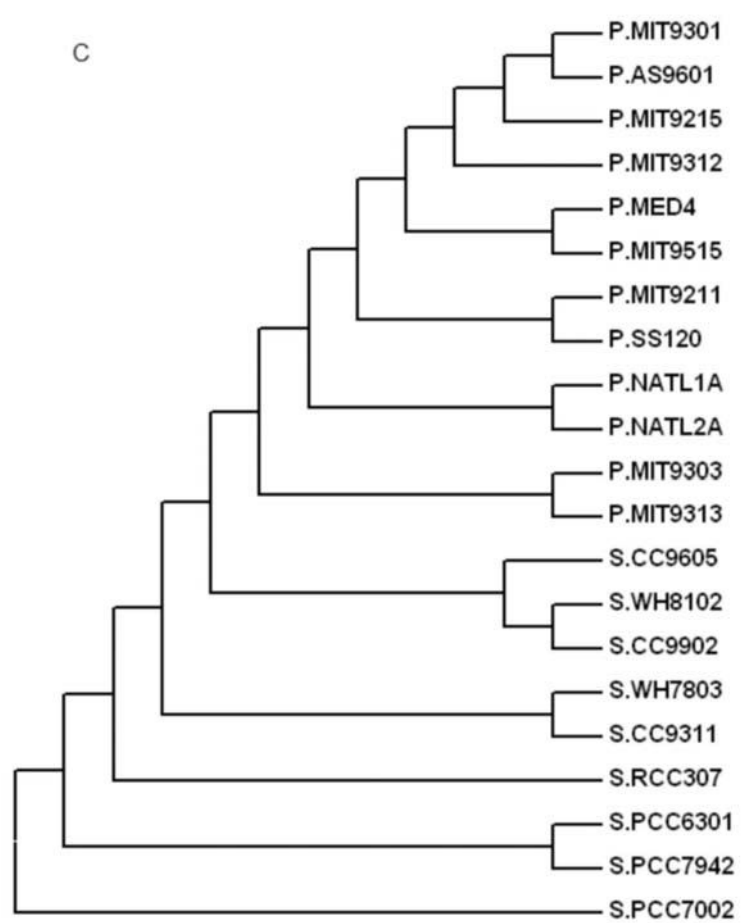
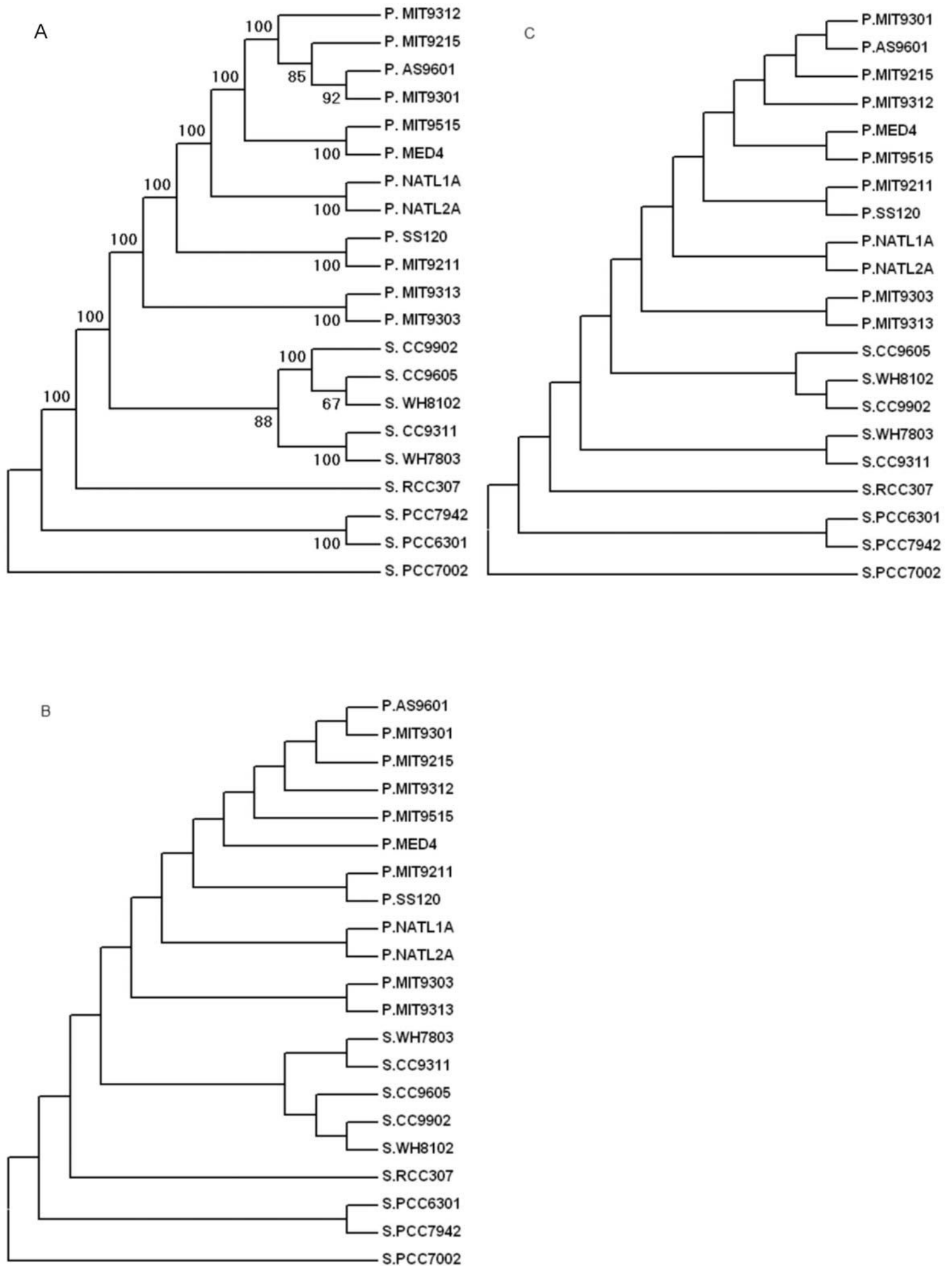


Figure 4