Maximum Likelihood Phylogenetic Reconstruction Using Gene Order Encodings

Fei Hu, Nan Gao, Meng Zhang and Jijun Tang

Abstract— **Gene order changes under rearrangement events such as inversions and transpositions have attracted increasing attention as a new type of data for phylogenetic analysis. Since these events are rare, they allow the reconstruction of evolutionary history far back in time. Many software have been developed for the inference of gene order phylogenies, including widely used maximum parsimony methods such as GRAPPA and MGR. However, these methods confronted great difficulties in dealing with emerging large nuclear genomes. In this study, we proposed three simple yet powerful maximum likelihood(ML) based methods for phylogenetic reconstruction by first encoding the gene orders into binary or multistate strings based on gene adjacency information presented in the given genomes and further converting these strings into molecular sequences. RAxML is at last used to compute the maximum likelihood phylogeny. We conducted extensive experiments using simulated datasets and found that although the multistate encoding is more complex and more time-consuming, it did not improve accuracy over the methods using simpler binary encodings. Among all methods tested in our experiments, MLBE is of the most accuracy in most cases and often returns phylogenies without errors. ML methods is also fast and in the most difficult case only takes up to three days to compute datasets with 40 genomes, making it very suitable for large scale analysis. We give three simple and robust phylogenetic reconstruction methods using different encodings based on maximum likelihood which has not been successfully applied for gene orderings before. Our development of these ML methods showed great potential in gene order analysis with respect to the high accuracy and stability, although formal mathematical and statistical analysis of these methods are much desired.**

I. Introduction

Genome rearrangements have been used to reconstruct deep evolutionary history because these gene rearrangements are "rare genomic events" [1]. Popular methods for gene order phylogeny include distance-based methods such as Neighbor-joining [2] and FastME [3], as well as parsimony methods such as GRAPPA [4] and MGR [5]. However, these methods all have their limitations. Distance-based methods are usually the fastest, however, they all rely on accurate computation of pairwise distances among genomes. When genomes are distant, these computation may severely under-estimate the true number of evolutionary events, making distance-based methods unreliable. On the other hand, GRAPPA and MGR 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 as the complexity of computing the median is essentially NP-hard.

Although recently great effort has been made to improve the accuracy of these methods, analyzing dozens of large nuclear genomes is still out of reach. To tackle the problem, we propose three ML based phylogenetic reconstruction methods using three different encodings: Maximum Likelihood on Binary Encoding(MLBE and MLBE2) and Maximum Likelihood on Multistate Encoding(MLME). In general gene orderings are transformed into either binary or multistate strings at first according to the encoding scheme and are further converted into molecular sequences(DNA or amino acid). Specifically MLBE converts the gene orderings into binary strings and uses amino acids to encode the strings into molecular sequences. MLBE2 is in essence similar to MLBE but uses nucleotides instead of amino acids, while MLME uses a more complex converting scheme to transfer gene orders into sequences. Finally RAxML [6] is utilized for the inference of phylogeny. We tested these methods on simulated datasets and found that MLBE has better accuracy against MLBE2 and is the best among all methods we tested. The remainder of the paper is organized as follows. In Section II we briefly review genome rearrangements and methods for rearrangement analysis, Section III describes binary encoding and the two methods based on it. Section IV describes the multistate encoding and the method MLME. We present our experimental results in Section V, using simulated datasets. Finally, we describe conclusions and future work in Section VI.

Fei Hu, Nan Gao and Jijun Tang are with the Department of Computer Science and Engineering, University of South Carolina, Columbia, USA (email: {hu5, gaon, jtang}@cse.sc.edu). Meng Zhang is with the College of Computer Science and Technology, Jilin University, China (email : zhangmeng@jlu.edu.cn)

II. Background

A. Gene order and genome rearrangements

Given a set of n genes $\{1, 2, \cdots, n\}$, a genome can be represented by an *ordering* of these genes. Each gene is assigned with an orientation that is either positive, written i, or negative, written $-i$. Two genes i and j are *adjacent* if i is immediately followed by j, or, equivalently, $-j$ is immediately followed by $-i$.

Let G be the genome with signed ordering of $1, 2, \cdots, n$. An *inversion* (also called *reversal*) between indices i and j $(i \leq j)$, produces the genome with linear ordering

$$1, 2, \cdots, i-1, -j, -(j-1), \cdots, -i, j+1, \cdots, n.$$

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

$$1, \cdots, i-1, j+1, \cdots, k, i, i+1, \cdots, j, k+1, \cdots, n.$$

978-1-4244-9895-6/11/$26.00 ©2011 IEEE

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

There are additional events for multiple-chromosome genomes, such as *translocation* (the end of one chromosome is broken and attached to the end of another chromosome), *fission* (one chromosome splits and becomes two) and *fusion* (two chromosomes combine to become one).

Given two genomes G_1 and G_2, we define the *edit distance* $d(G_1, G_2)$ 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 [7] 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. [8] later showed that this edit distance can be computed in linear time.

Yancopoulos et al. [9] proposed a universal double-cut-and-join (DCJ) operation that accounts for common events such as 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.

B. Methods for gene order phylogenetic reconstruction

There exist various methods for reconstructing the gene order phylogenies, including distance-based methods (neighbor-joining (NJ) [2] and FastME [3]), Bayesian method (Badger [10]) and maximum parsimony methods (GRAPPA [4] and MGR [5]).

Neighbor-joining and its variants (including FastME) are the most popular distance-based methods which run in polynomial time and have shown great performance both in synthetic and biological gene order data [11]. These methods all require the input of a matrix of pairwise evolutionary distances, such as inversion distance or DCJ distance. Since NJ and FastME do not have an optimization criteria to follow, the accuracy of the reconstructed phylogeny requires the estimated pairwise distance to be as close to the true distance as possible. When genomes are distant, inversion and DCJ distances will severely under-estimate the true number of events, hence some form of corrections are needed. *Empirical derived estimation* [12] estimates the true number of inversions in which the minimum number of inversions is initially computed between two genomes and an empirical correction is applied based on a statistical model to estimate the true inversion distance. Another more recent approach relies on a structural characterization of a genome pair under the DCJ model [9]. This method (called CDCJ in this study) successfully postpones the emergence of *saturation* [13].

Within the maximum parsimony (MP) framework, GRAPPA and MGR explicitly search the best tree with the minimum number of evolutionary events and are generally more accurate than distance-based methods. However, using GRAPPA and MGR to compute phylogeny for organismal genomes with many events is extremely expensive, because their computation takes time exponential in both the size of the genomes and the distances among genomes. As a result, only relatively small dataset (such as a dozen organelle genomes) can be handled.

Several other methods have been proposed. For example, MPBE (Maximum Parsimony on Binary Encoding) [14] transforms adjacency pairs from the signed permutation into sequence-like strings. These strings are then converted into DNA sequences and computed using ordinary sequence parsimony software (e.g. PAUP* 4.0 [15]) to obtain a phylogeny. Wang et al. later introduced a new set of encodings called MPME (Maximum Parsimony on Multistate Encoding) [16] to improve the accuracy. These encoding based parsimony methods possess slightly better accuracy compared to the neighbor-joining method, yet they are computationally very expensive.

Maximum likelihood (ML) is often considered the best approach in sequence phylogeny analysis [17]. In this approach, each tree is assigned a likelihood based on all possible ancestral sequences. Among all possible tree topologies, the one with the highest likelihood is chosen as the phylogeny. Not long ago ML approach was not widely used due to its poor scalability. As a result, the binary and multistate encoding approaches have been only tested with maximum parsimony methods, although in principle these encodings could be applied to the ML framework.

Recent algorithm developments and the introduction of high-performance computation tools such as RAxML [6] have made the ML approach feasible for large scale analysis of molecular sequences. These improvements motivated us to develop the ML method for gene order phylogeny analysis. Our simulation results showed that ML methods are very accurate and maintain very high accuracy for difficult datasets of abundant genomes and high events per edge when all the existing methods fail. Our experiments also showed that although using similar encodings, using maximum likelihood approaches have brought much better accuracy than using maximum parsimonies.

III. Maximum likelihood methods based on binary encoding

In this section, we will first describe MLBE in detail and then introduce MLBE2 which uses another scheme to code the binary strings.

A. MLBE

Let G be a signed permutation of n genes. For linear genomes, genes 0 and $n+1$ are added to indicate the start and end of a genome respectively. For the pair $(i, j), 0 \leq i, j \leq n+1$, we set up a character to indicate the presence or absence of this adjacency. If i is immediately followed by j in the gene ordering, or $-j$ is immediately followed by $-i$, 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 $\binom{2n+2}{2}$ possible adjacencies, we can

$$G_1 \quad : \quad (4, 5, -2, -1, -3) \qquad (1)$$
$$G_2 \quad : \quad (-1, 4, 2, -3, -5) \qquad (2)$$
$$G_3 \quad : \quad (3, 2, -5, -4, -1) \qquad (3)$$

(a) Three signed linear genomes

TABLE I

EXAMPLE OF THE BINARY ENCODING (0 INDICATES THE START OF A GENOME, 6 INDICATES THE END OF A GENOME).

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

(b) Binary Encoding

further reduce the length of these sequences by removing those characters at which every genome has the same state. Table III-A 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. After converting the gene orders into strings of 0 and 1, we further convert these sequences into amino acid sequences and utilize the power of those widely used ML packages developed for molecular sequences. We tested several ML packages such as TREE-PUZZLE [18] and GARLI [19] and among them, RAxML [6] is the best by incorporating the rapid bootstrapping [20]. In the method of MLBE, the following steps are used to convert 1 and 0 into amino acids:

- For a dataset, randomly pick an amino acid to code *absent state*.
- For an adjacency, code *present state* by randomly picking one from the remaining 19 amino acids, such choice will be preserved for all genomes on the site corresponding to the given adjacency.

Table III-A shows an example of amino acid sequences produced by MLBE from the binary strings of the genome presented in Table III-A.

B. MLBE2

MLBE2 uses a simple code that treats every 1 as A or T, and consequently every 0 as C or G. Like what we regulate in MLBE, at the site of a given adjacency across all genomes, we enforce the presence of nucleotides to be either A associating with C or T associating with G. Such choice is randomly assigned so that a balanced number of the two pairings is presented in the new sequences.

Table III-B shows the example of nucleotide coding of the binary strings of the genomes presented in Table III-A. Again, RAxML will be used to obtain trees from these nucleotides sequences.

IV. MAXIMUM LIKELIHOOD METHOD BASED ON MULTISTATE ENCODING

Bryant [21] proposed an encoding method called Multistate Encoding. Let n be the number of genes in each genome; then each gene order is translated into a sequence with $2n$ characters. For every gene i, $1 \leq i \leq n$, site i takes the value of the gene immediately following i; site $n+i$ takes the gene immediately following gene $-i$. Table IV shows an example of such encoding.

Once the states of each site are determined for all genomes, we can easily convert them into molecular sequences by randomly assigning amino acid to a given state and then use RAxML to compute the phylogeny. We call such method MLME in this study and an example of converted amino acid sequences is shown in Table IV. Since RAxML only deals with 20 amino acids, hence no site can have more than 20 states. As a result, MLME is limited to handle datasets with no more than 20 genomes at this stage.

V. EXPERIMENTAL RESULTS

We tested these ML methods (MLBE, MLBE2 and MLME) using simulated datasets. In our simulations, we generated model tree topologies from the uniform distribution on binary trees, each with 10, 20 and 40 leaves. We chose genomes of 200 and 1,000 genes, spanning the range from organelles to small bacteria. On each tree, we evolved signed permutations using various numbers of evolutionary rates: letting r denote the expected number of rearrangement events (80% inversion and 20% transposition) along an edge of the true tree, we used values of $r = 20, 35, \cdots, 80$ for 200 genes and $r = 100, 175, \cdots, 400$ for 1000 genes. The actual number of events along each edge was sampled from a uniform distribution on the set $\{1, 2, \cdots, 2r\}$. For each combination of parameter settings, we ran 10 datasets and averaged the results.

We compared ML methods with other two methods: FastME with the true distance estimator based on DCJ distance (CDCJ) [13], FastME with Empirical Distance Estimation (EDE) [12]. MPBE was also added to the test only with the datasets of 10 genomes so that MPBE can accomplish the test with branch-and-bound search in an appropriate time. Since neither GRAPPA nor MGR can finish any of the above tests within days of computation, we therefore conducted a special experiment to accommodate GRAPPA and MGR by simulating datasets of 10 genomes matching mitochondrial DNA consisting of 37 genes where transpositions are dominant [22]. In particular, the GRAPPA was configured to

TABLE II

EXAMPLE OF THE CONVERTED SEQUENCES USING MLBE, V IS PICKED TO ENCODE ABSENT STATE IN ALL SEQUENCES.

	Adjacencies															
	H,4	4,5	5,-2	-2,-1	-1,-3	-3,T	H,-1	-1,4	4,2	2,-3	-3,-5	-5,T	H,3	3,2	-4,-1	-1,T
G_1	Q	K	S	A	N	A	V	V	V	V	V	V	V	V	V	V
G_2	V	V	V	V	V	V	W	Q	R	C	Y	Y	V	V	V	V
G_3	V	K	S	V	V	V	V	V	V	V	V	V	L	F	M	H

TABLE III

EXAMPLE OF THE CONVERTED SEQUENCES USING MLBE2.

	Adjacencies															
	0,4	4,5	5,-2	-2,-1	-1,-3	-3,6	0,-1	-1,4	4,2	2,-3	-3,-5	-5,6	0,3	3,2	-4,-1	-1,6
G_1	T	T	A	T	T	T	C	G	G	C	C	C	C	G	G	C
G_2	G	G	C	G	G	G	A	T	T	A	A	A	C	G	G	C
G_3	G	T	A	G	G	G	C	G	G	C	C	C	A	T	T	A

$$G_1 \quad : \quad (4, 5, -2, -1, -3) \tag{4}$$
$$G_2 \quad : \quad (-1, 4, 2, -3, -5) \tag{5}$$
$$G_3 \quad : \quad (3, 2, -5, -4, -1) \tag{6}$$

(a) Three signed linear genomes

TABLE IV

EXAMPLES OF MULTISTATE ENCODING, 0 INDICATES THE START OF A GENOME, 6 INDICATES THE END OF A GENOME.

	genes									
	1	2	3	4	5	−1	−2	−3	−4	−5
G_1	2	−5	1	5	−2	−3	−1	6	0	−4
G_2	0	−3	−2	2	3	4	−4	−5	1	6
G_3	4	−5	2	5	−2	6	−3	0	−1	−4

(b) Multistate Encoding

TABLE V

EXAMPLE OF THE CONVERTED SEQUENCES USING MLME.

	genes									
	1	2	3	4	5	−1	−2	−3	−4	−5
G_1	G	D	R	L	Y	W	N	S	K	E
G_2	N	K	M	T	C	P	M	G	W	V
G_3	C	D	P	L	Y	Q	I	H	V	E

use Caprara's inversion median solver [23] and enable EDE distance estimator. And MGR was tested given the parameter $-c$ and $-H1$ for efficiency and speed. Similarly the number of events was the value of $r = 2, 3, \cdots, 6$ and the actual events were sampled from the set of $\{1, 2, \cdots, 2r\}$. Finally we ran RAxML with the same setting but on binary strings as a control test to demonstrate the efficiency of our three approaches of encodings which is called RAxML-Binary in this study.

A. Topological Accuracy

We assess topological correctness by computing the *false negatives* (FN) and *false positives* (FP) [24] rates. The *false negatives* are those edges in the true tree but not in the inferred tree. The *false positives* are those edges in the inferred tree that do not exist in the true tree. The *false negatives rate* is the number of false negatives divided by the number of internal edges. The *false positives rate* is similarly defined. The *Robinson-Foulds* (RF) rate is then defined as the average of the FN and FP rates. An RF rate of more than 5% is generally considered too high [25].

Figure 1 and Figure 2 show the topological accuracy of these methods (MLME is not applicable for 40-genome datasets and MPBE is too slow to finish the brand-and-bound search for datasets containing 20 and 40 genomes). Both figures show that MLBE was of the most accuracy in most of the cases when genome number is 20 and 40, except for a few occasions when MLME becomes the best(20 genomes, 200 genes, fewer than 50 events). As to the results for datasets of 10 genomes, MLBE and distance methods quite matched each other in performance and both outperformed the other methods. Figure 3 shows the results of simulated mitochondrial gene orderings with only transpositions applied when GRAPPA and MGR were also present in the contest. The results suggested the MLBE possessed the greatest performance in all conditions

TABLE VI

TIME USAGE OF ML METHODS ON 200 GENES(- INDICATES MISSING DATA SINCE MLME CANNOT BE USED FOR MORE THAN 20 GENOMES).

methods	time (in minutes)														
	r=20			r=35			r=50			r=65			r=80		
	N=10	N=20	N=40	N=10	N=20	N=40	N=10	N=20	N=40	N=10	N=20	N=40	N=10	N=20	N=40
MLBE	6	30	222	7	48	270	10	96	318	12	108	426	15	150	468
MLBE2	0.3	2	7	0.5	4	9	1	7	12	1.5	9	25	2	12	34
MLME	18	72	–	30	84	–	35	144	–	40	516	–	65	948	–

TABLE VII

TIME USAGE OF ML METHODS ON 1000 GENES (- INDICATES MISSING DATA SINCE MLME CANNOT BE USED FOR MORE THAN 20 GENOMES).

methods	time (in minutes)														
	r=100			r=175			r=250			r=325			r=400		
	N=10	N=20	N=40	N=10	N=20	N=40	N=10	N=20	N=40	N=10	N=20	N=40	N=10	N=20	N=40
MLBE	18	132	354	20	138	378	24	174	414	28	222	708	30	240	756
MLBE2	1	4	15	1.5	6	25	2	10	30	2.6	12	43	2.8	15	55
MLME	85	324	–	90	408	–	110	558	–	205	1590	–	400	4200	–

compared to the parsimony methods(GRAPPA, MGR and MPBE) and distance methods(FastME-CDCJ and FastME-EDE). In contrast RAxML-Binary is significantly worse in accuracy and stability in most of cases. Although MLBE and MLBE2 are both based on the same principle of binary encoding, MLBE is more accurate by using amino acids to code the binary strings. The performance of ML methods improves with more genes, indicating that the length of the sequences has big impact on their accuracy.

FastME was always the fastest in our testing, while the speed of these ML methods were acceptable. Table V and V present the average time used by these methods.

These two tables show that MLBE2 is very fast and generally requires less than one hour to compute, while MLME is very slow and may take up to three days to finish. MLBE is much slower than MLBE2 and its speed quickly decreases with the increase of number of characters. However, it only requires fewer than 13 hours even for the most difficult datasets. Comparing to the results in Figures 1 and 2, such computation time is worthwhile and is easily offset by the increased accuracy of inferred phylogenies. Since all our tests were conducted on single processors and did not use the parallel version of RAxML, MLBE has the potential to handle several dozens of large nuclear genomes if the full computational power of RAxML is utilized.

VI. CONCLUSION

In this paper, we presented a set of three maximum likelihood methods (MLBE, MLBE2 and MLME) for gene order phylogenetic reconstructions. Our tests on simulated datasets show that these methods are very accurate and scale well to accommodate emerging large genomes. Our tests also suggest that MLBE is the most accurate and generally outperforms all methods we have tested. Since the encoding of MLBE is based on the presence and absence of gene adjacencies, it will be relatively easy to extend MLBE to handle other events such as gene loss and gene duplication. Our development of these maximum likelihood methods showed great potential of ML approaches in gene order analysis, and formal mathematical and statistical analysis of these methods are much desired.

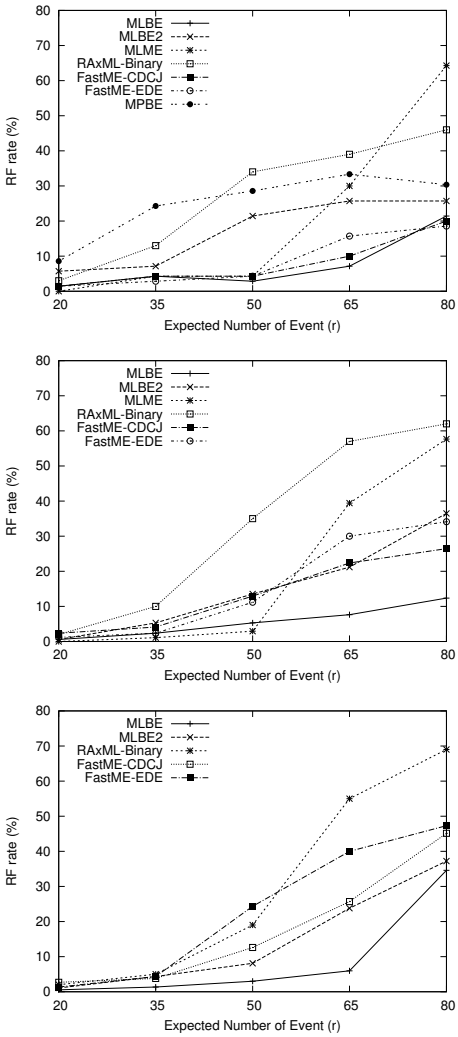

Fig. 1. RF rates for MLBE, MLBE2, MLME, RAxML-binary, FastME-CDCJ, FastME-EDE, MPBE on 200-gene datasets (top: 10 genomes, middle: 20 genomes(MPBE was excluded), bottom: 40 genomes(MLME and MPBE were excluded)).

ACKNOWLEDGEMENTS

FH, NG, JT are partly supported by grants NSF 0904179 and NIH GM078991-03S1. ZM is supported by China's

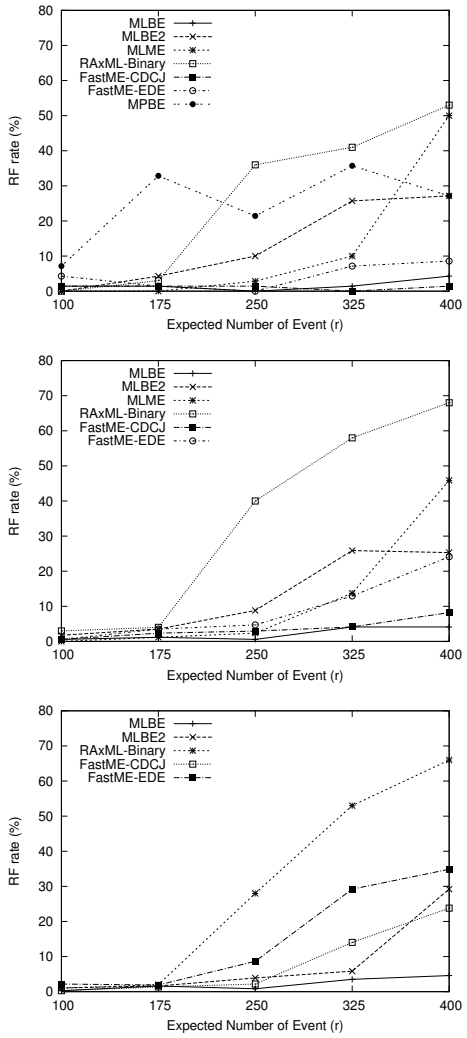

Fig. 2. RF rates for MLBE, MLBE2, MLME, RAxML-binary, FastME-CDCJ, FastME-EDE, MPBE on 1000-gene datasets (top: 10 genomes, middle: 20 genomes(MPBE was excluded), bottom: 40 genomes(MLME and MPBE were excluded)).

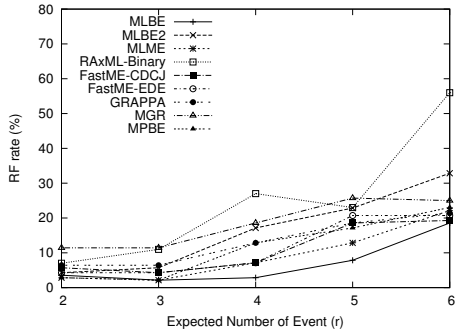

Fig. 3. RF rates for MLBE, MLBE2, MLME, RAxML-binary, FastME-CDCJ, FastME-EDE, GRAPPA, MGR, MPBE on simulated 37-gene and 10-genome datasets where only transposition existed.

Fundamental Research Funds for the Central Universities (No.200903186). All experiments were conducted on a 128-core shared memory computer supported by US National Science Foundation grant (NSF grant number CNS 0708391).

REFERENCES

[1] Rokas A, Holland P: Rare genomic changes as a tool for phylogenetics. Trends in Ecology and Evolution 2000, 15:454–459.

[2] Saitou N, Nei M: The neighbor-joining method: A new method for reconstructing phylogenetic trees. Mol. Biol. Evol. 1987, 4:406–425.

[3] Desper R, Gascuel O: Fast and accurate phylogeny reconstruction algorithms based on the minimum evolution principle. J. Comput. Biol. 2002, 9:687–705.

[4] 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.

[5] Bourque G, Pevzner P: Genome-scale evolution: reconstructing gene orders in the ancestral species. Genome Research 2002, 12:26–36.

[6] Stamatakis A: RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. bioinformatics 2006, 22:2688-2690.

[7] 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.

[8] 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.

[9] Yancopoulos S, Attie O, Friedberg R: Efficient sorting of genomic permutations by translocation, inversion and block interchange. Bioinformatics 2005, 21:3340–3346.

[10] Larget B, Kadane J, Simon D: A Bayesian approach to the estimation of ancestral genome arrangements. Mol. Phy. Evol. 2005, 36:214–223.

[11] Wang L, Jansen R, Moret B, Raubeson L, Warnow T: Distance-based genome rearrangement phylogeny. J. Mol. Evol. 2006, 63:473–483.

[12] Moret B, Wang L, Warnow T, Wyman S: New approaches for reconstructing phylogenies from gene order data. Bioinformatics 2001, 17:S165–S173.

[13] Lin Y, Moret B: Estimating true evolutionary distances under the DCJ model. Bioinformatics 2008, 24:i114–i122.

[14] 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.

[15] Swofford D: PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4. Sunderland, MA 2003.

[16] 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.

[17] Felsenstein J: Evolutionary trees from DNA sequences: a maximum-likelihood approach. Journal of molecular evolution 1981, 17:368–376.

[18] Schmidt Heiko A., Strimmer Korbinian, Vingron Martin, von Haeseler Arndt: TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. bioinformatics 2002, 18:502–504.

[19] Zwickl D. J., Strimmer Korbinian, Vingron Martin, von Haeseler Arndt: Genetic algorithm approaches for the phylogenetic analysis of large biological sequence. Ph.D. dissertation, The University of Texas at Austin.

[20] Stamatakis A, Hoover P, Rougemont J: A rapid bootstrap algorithm for the RAxML web-servers. Syst. Biol. 2008, 75:758–771.

[21] Bryant D: A lower bound for the breakpoint phylogeny problem. In Proc. 11th Ann. Symp. Combinatorial Pattern Matching CPM'00. Edited by Giancarlo R, Sankoff D, Springer 2000:235–247.

[22] Sorenson MD, Fleischer RC: Multiple independent transpositions of mitochondrial DNA control region sequences to the nucleus. In Proc. National Academy of Sciences USA, Volume 93 1996:15239–15243.

[23] A C: The Reversal Median Problem. INFORMS Journal on Computing 2003, 15(1):93–113.

[24] Robinson D, Foulds L: Comparison of phylogenetic trees. Mathematical Biosciences 1981, 53:131–147.

[25] Swofford D, Olson G, Waddell P, Hillis D: Phylogenetic inferences. In Molecular Systematics, 2nd edition. Edited by Hillis DM, Moritz C, Mable B 1996.