

Isolating - A New Resampling Method for Gene Order Data

Jian Shi, William Arndt, Fei Hu and Jijun Tang

Abstract—The purpose of using resampling methods on phylogenetic data is to estimate the confidence value of branches. In recent years, bootstrapping and jackknifing are the two most popular resampling schemes which are widely used in biological research. However, for gene order data, traditional bootstrap procedures can not be applied because gene order data is viewed as one character with various states. Experience in the biological community has shown that jackknifing is a useful means of determining the confidence value of a gene order phylogeny. When genomes are distant, however, applying jackknifing tends to give low confidence values to many valid branches, causing them to be mistakenly removed. In this paper, we propose a new method that overcomes this disadvantage of jackknifing and achieves better accuracy and confidence values for gene order data. Compared to jackknifing, our experimental results show that the proposed method can produce phylogenies with lower error rates and much stronger support for good branches. We also establish a theoretic lower bound regarding how many genes should be isolated, which is confirmed empirically.

I. BACKGROUND

A phylogeny is a representation of the evolutionary history of a collection of organisms or genes, known as taxa. A phylogenetic reconstruction is usually depicted as a tree, in which modern taxa are at the leaves and ancestral taxa are represented by internal nodes. The edges of the tree denote the evolutionary relationships among the various taxa. Although DNA or protein sequences are still the primary source of data for phylogenetic analysis, genome rearrangements have been used to reconstruct deep evolutionary history because these rearrangements are “rare genomic events” [1].

For sequence data, biologists normally use a bootstrap procedure to estimate the quality of phylogenetic trees by assigning confidence values to their edges [2]. Generally, edges with high confidence values ($> 75 - 80\%$) are considered reliable. However, bootstrapping does not suit gene order data because in parsimony terms a gene order data set is a single character with a very large number of potential states [3].

Jackknifing has been used to assess the quality of gene order phylogenies [4], [5] and its performance has been systematically studied by Shi et al. [6]. Through extensive experiments, it was suggested up to 40% of genes can be removed and 85% is a good threshold of confidence values. The weakness of jackknifing is its inability to produce reliable results when the data set contains distant genomes: a high portion ($> 50\%$) of branches are mistakenly discarded yielding a highly unresolved tree.

Jian Shi, William Arndt, Fei Hu and Jijun Tang are with the Department of Computer Science and Engineering, University of South Carolina, Columbia, USA (email: {shi2, arndtw, hu5, jtang}@cse.sc.edu).

In this paper, we propose a new resampling method for gene order data called *isolating* which will reduce those errors. Moreover, we analyze the theoretical mechanism of this isolating method and produce a simple formula to determine the amount of genes which should be isolated from the original dataset. This value is confirmed by our experiments.

A. Gene Orders and Rearrangements

For n genes $\{g_1, g_2, \dots, g_n\}$, a genome can be represented by an *ordering* of these genes. Each gene is assigned with an orientation that is either positive, written in g_i , or negative, written in $-g_i$. Two genes g_i and g_j are *adjacent* if g_i is immediately followed by g_j , or, equivalently, $-g_j$ is immediately followed by $-g_i$.

Gene orders can be rearranged through events such as inversions and transpositions. Let G be the genome with signed ordering of g_1, g_2, \dots, g_n . An *inversion* between indices i and j ($i \leq j$) produces the genome with linear ordering

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

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

The double-cut-and-join (DCJ) model was initially proposed by Yancopoulos et al. [7]. Given two consecutive genes a and b , depending on their respective orientation, their adjacency can be of four different types: $\{a_h, b_l\}, \{a_h, b_h\}, \{a_l, b_l\}, \{a_l, b_h\}$. A DCJ operation makes a pair of cuts and reconnects the telomeres created by those cuts in a new arrangement. Thus DCJ can mimic rearrangement events such as inversion, fission, fusion, translocation and transposition through different combinations of one or more DCJ operations.

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 genome into the other. When only inversions are allowed, the edit distance is the *inversion distance*. Hannenhalli and Pevzner [8] developed a mathematical and computational framework for generating the inversion distance between two signed gene-orders in polynomial time.

Edit distance does not reflect the number of events which occurred in the true evolutionary history; sometimes events can erase the evidence of previous events which causes the true distance to be underestimated. Several distance correction methods have been proposed, including the EDE correction

by Wang et al. [9] and the CDCJ correction by Lin and Moret [10].

In the CDCJ model, Lin and Moret considered the four possible cases of changes on adjacencies and telomeres, and derived a novel process to estimate the distance between two genomes G_1 and G_2 by computing every intermediate state, step by step from G_1 to G_2 , until they either match or reach to a pre-defined threshold. In this manner CDCJ can estimate the actual number of DCJ operations that took place in the evolutionary history. According to simulations, CDCJ produces more accurate pair-wise genome distances than traditional DCJ.

There are several widely used methods to reconstruct phylogenies from gene order data which include distance-based methods (neighbor-joining [11] and FastME [12]), Bayesian (Badger [13]) and direct optimization methods (GRAPPA [14] and MGR [15]). Although Badger, GRAPPA and MGR are more accurate, these methods are very demanding computationally and may not be able to analyze data sets with distant genomes. NJ or FastME accept any metric distance and simulations have shown that using corrected distance (such as EDE and CDCJ) produces far more accurate phylogenies than using edit distances.

B. Resampling Methods

Resampling methods are widely used in phylogenetic reconstruction to place confidence values on inferred phylogenies. A well known resampling scheme is bootstrapping, which is widely accepted in the phylogenetic reconstruction of sequence data sets. In general, bootstrapping works by sampling columns with replacement and collecting them until a new data set the same size as the original has been created. Each of the replicates is then analyzed and a new phylogeny is inferred from it. At the end, a consensus tree is constructed which gathers all the inferred monophyletic groups that occurred in a majority of the bootstrap replicates [2]. Confidence values are assigned to branches based on the frequency that a given branch appears in all the replicate trees.

Due to the nature of the bootstrap procedure, sampling columns with replacement, it can not be applied to gene order data because a gene order genome is composed of a single column with a large number of potential states [3]. For this reason, the adaptation of bootstrapping to gene order data is highly impractical.

Another resampling method called jackknifing has been adapted to assess gene order phylogenies. The process of jackknifing is as follows:

- 1) Randomly choose $m\%$ genes in the data set.
- 2) Remove chosen genes from all genomes in the data set.
- 3) Compute a phylogeny from the new data set with fewer genes.
- 4) Repeat steps 1-3 k times to obtain k trees
- 5) Compute a strict consensus from these k trees, as well as confidence values on the branches.

Shi et al. thoroughly studied the performance of jackknifing for gene order data and proved that jackknifing is a useful

means of determining the confidence level of a phylogeny [6]. The percentage of genes to be removed is best set at $m = 40$ which contrasts with most existing literature that uses $m = 50$. It has also been found that the number of replicates (k) varies among datasets, although most require fewer than 100. On the other hand, results on certain data sets show that jackknifing has the unwelcome tendency of pruning up to 50% of valid branches.

II. SIMULATION STRATEGY

We performed large groups of simulations to determine the quality of our proposed resampling method as access to the true evolutionary relationships is needed to judge performance. In this study, we have generated model tree topologies from the uniform distribution of binary trees, each with 20 or 40 leaves. The number of genes on each leaf is set as 100.

Let r denote the expected number of inversions along an edge of the true tree, we use values of $r = 4, 8, 12, \dots, 32$. The actual number of inversion events along each edge is sampled from a uniform distribution on the set $\{r/2, r/2 + 1, \dots, 3r/2\}$. For each combination of parameter settings, we generated 100 datasets and averaged the results. We conducted experiments with 200 genes and obtained similar results. Due to space limitations those results are not shown here.

FastME was chosen to obtain phylogenies based on its speed and accuracy [12]. Other methods (GRAPPA and MGR) generally cannot generate results in a practical time frame for any tree in this study and are thus excluded. Any distance can be used in FastME, among them we found that the corrected DCJ model proposed by Lin and Moret [10] fits well in our experiments. The EDE distance proposed by Wang et al. was used in some test cases as well.

The quality of inferred phylogenies have been assessed using the Robinson and Foulds (RF) rates [16]. Assuming T be the true tree and T' be the inferred tree. If an edge e in T but not in T' , e is reported as a false negative (FN). The false negative rate is the number of false negative edges divided by the number of internal edges. The false positive (FP) is defined similarly by swapping T and T' . The RF rate is defined as the average of the FN and FP rates.

The bootstrapping and jackknifing procedures produce a consensus tree with confidence values assigned to all branches. As discussed earlier, a branch with confidence value below 85% will be discarded. Discarded edges are shown in the consensus tree by merging the two nodes linked by the pruned edge into one. As a result of this the FN rates are increased. The goal of a resampling method is to remove FP edges by contracting low support branches while still retaining low FN rates. Our previous study suggests that jackknifing does not meet the goal: although it generally can provide phylogenies with very low (close to 0%) FP edges, the FN rates are too high for distant genomes.

III. NEW RESAMPLING SCHEMES

To overcome the problems of jackknifing, we have explored several different approaches. These approaches are based on

G_1	1	2	3	4	5	6	7\$
G_2	4	3	1	2	-7	-6	5\$
G_3	1	4	3	5	-6	7	2\$
		↓					
G_1	1	4	5	6	7\$	2	3\$
G_2	4	2	-7	-6	5\$	3	1\$
G_3	1	5	-6	7	2\$	4	3\$

Fig. 1. Isolating a continuous segment. (top) Original genomes viewed as columns, and columns 2 and 3 are picked for isolation. (bottom) New genomes with new chromosomes. \$ indicates the end of a chromosome.

a procedure we refer to as *isolation*. A gene is isolated by applying a DCJ event to its current chromosome that removes the gene and places it in a new circular chromosome. This approach has several advantages when compared to jackknifing procedures. The isolated genes are still present in the data set, so there is no worry about creating data sets with unequal gene content.

Compared to jackknife sampling, when a gene is chosen, it must be removed from every genome in the data set which limits the number of random choices to the number of genes. In contrast, individual genes can be isolated without forcing any choices elsewhere which allows a number of random choices equal to the number of genomes multiplied by the number of genes. Isolation also makes sense when considering the goal of resampling is to randomly remove information from the data set. Once a gene has been isolated, a single DCJ event can reinsert that gene at any location or orientation in the entire genome; effectively no information remains about the proper location of that gene.

A. Isolating A Continuous Segment—First Attempt

The first approach we examined was to simulate sequence resampling by viewing the gene orders as columns, randomly choosing a sequence of columns, and isolating these sequences into their own circular chromosomes. This process is shown in Fig. 1. This process is repeated 100 times and a consensus tree with confidence values is constructed. To test the performance of this approach and to seek the optimal segment length, we conducted simulations on 20 genomes, using segment lengths from 5 to 30. All branches with confidence values below 85% are contracted, and the FP and FN rates are shown in Fig. 2.

This figure clearly shows that segment lengths have very little impact on the consensus tree, even though the lengths are very different. It also shows that the FN and FP rates of the consensus tree are similar to the rates of the tree obtained from the original genomes. In other words, it fails to identify FP edges. These findings suggest that removing only one segment does not introduce enough disturbance.

To explain this conclusion we can examine the changes from the point of view of gene adjacencies. Given a sample set of n genes $\{g_1, g_2, \dots, g_i, \dots, g_j, \dots, g_n\}$, this approach picks segment g_i to g_j and makes it a new chromosome after g_n . During this process, two adjacencies $\{g_{i-1}, g_i\}, \{g_j, g_{j+1}\}$

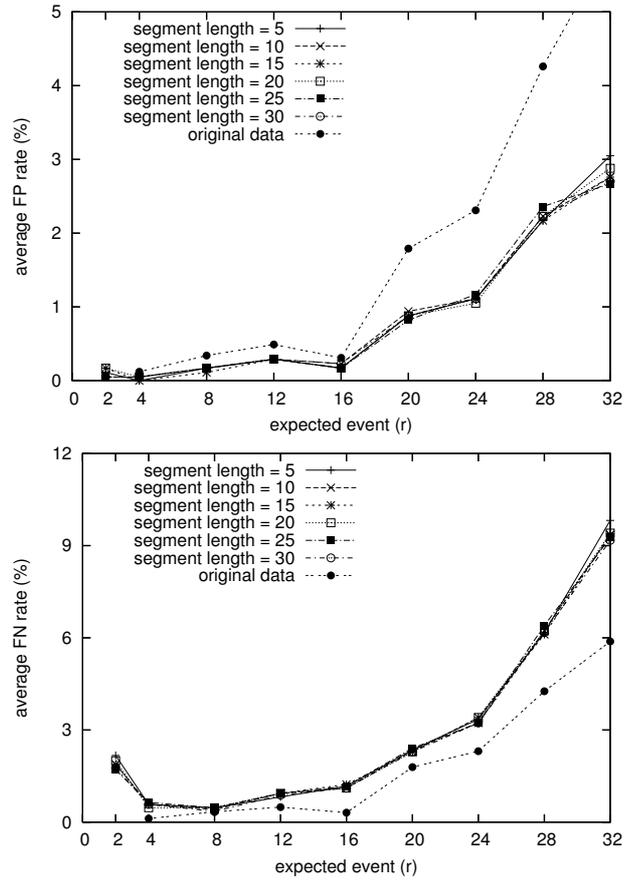


Fig. 2. upper: FP rates of different segment lengths; lower: FN rates of different segment lengths

are broken and new adjacency $\{g_{i-1}, g_{j+1}\}$ is made, while all other adjacencies remain unchanged. No matter how we change the number of genes in the segment, the change of adjacencies are the same. Thus the disturbance to the pairwise distances is small and similar among the replicates. To obtain a better resampling method, we need to isolate more genes from different locations, and thus disturb more adjacencies.

B. Isolating Genes from Multiple Locations

The segment experiment shows that more adjacencies should be disturbed to obtain a good assessment. We explored the following approaches (Fig. 3):

- 1) (Fig. 3 a): Randomly pick m single columns from the dataset and attach these columns as single-gene chromosomes. As indicated by previous subsection, the length of segments is not important, thus picking single-gene columns will have similar effect as picking multi-gene columns.
- 2) (Fig. 3 b): Randomly pick the same set of m genes and attach these genes as single-gene chromosomes. This is similar to the jackknife approach, except that these m genes are not discarded but retained in the genome in extra chromosomes.

G_1	1	2	3	4	5	6	7\$	→	G_1	1	3	5	6	7\$	2\$	4\$
G_2	4	3	1	2	-7	-6	5\$		G_2	4	1	-7	-6	5\$	3\$	2\$
G_3	1	4	3	-5	-6	7	2\$		G_3	1	3	-6	7	2\$	4\$	-5\$

(a)

G_1	1	2	3	4	5	6	7\$	→	G_1	1	3	4	6	7\$	2\$	5\$
G_2	4	3	1	2	-7	-6	5\$		G_2	4	3	1	-7	-6\$	2\$	5\$
G_3	1	4	3	-5	-6	7	2\$		G_3	1	4	3	-6	7\$	2\$	-5\$

(b)

G_1	1	2	3	4	5	6	7\$	→	G_1	1	3	4	6	7\$	2\$	5\$
G_2	4	3	1	2	-7	-6	5\$		G_2	4	3	2	-6	5\$	1\$	-7\$
G_3	1	4	3	-5	-6	7	2\$		G_3	1	3	-5	7	2\$	4\$	-6\$

(c)

Fig. 3. Isolating genes from multiple locations, \$ indicates the end of a chromosome. (a) Isolating multiple single columns. (b) Isolating the same set of genes from all genomes (genes circled in the left are isolated). (c) Isolating different set of genes from the genomes.

- 3) (Fig. 3 c): Randomly pick m genes from a genome and attach these genes as single-gene chromosomes; repeat this process for all genomes.

A potential pitfall exists when the same gene is isolated in multiple genomes. A single gene appearing on a circular chromosome in multiple genomes will cause the DCJ distance between them to decrease; for this reason isolating the same genes from two genomes will cause long branch attraction and distort the consensus tree.

Assuming we have N genomes and each has n genes, we have n choices to isolate one gene from the first genome, $n - 1$ choices from the second genome and $n - N$ choices from the last genome. Under ideal circumstances, we can isolate each genome n/N times while keeping the probability of co-isolating genes under control. As a result, the target value of isolated genes for each genome is:

$$m = \lceil \frac{n}{N} \rceil + 1 \quad (1)$$

We conducted simulations to test the performance of these options by picking $m = 6$ (using 100 genes and 20 genomes). Fig. 4 shows the results by contracting edges with confidence values below 85%. From this figure, we observe that the FP and FN rates for the first and third options are almost identical throughout the test, while the second option presents much higher errors.

Randomly picking columns or individual genes introduces similar amount of disturbance to gene adjacencies, hence similar disturbance to the pairwise distances, making the first and third option almost the same. From this experiment, we chose the first option (isolating multiple single-gene columns) as our new resampling method.

IV. EXPERIMENTAL RESULTS FROM THE ISOLATING METHOD

From the above experiment, we chose the first option (isolating multiple single-gene columns) as our new resampling

method and tested its performance using simulated datasets with 100 genes.

A. Topological Accuracy

We first examined the topological accuracy of the new resampling method by using various number of isolated genes and compared the results with jackknifing. Figs. 5 and 6 show the FP and FN rates by contracting low support edges. From these figures, the best choice for the number of genes to isolate is indeed the value determined by Eq. 1 (6 times for 20 genomes, 4 times for 40 genomes)—more amount of isolating may introduce too much disturbance, resulting in high FN rates.

Compared to the results of jackknife, the FP rates are similarly very low ($< 1\%$), while the FN rates are much improved (about 50% reduction), making trees more resolved.

B. Assessment of Confidence Values

Shi et al. suggested 85% is a near optimal threshold for the confidence values. Branches with higher confidence values can be treated as correct [6]. However based on their findings, almost two thirds of valid branches are mistakenly thrown out due to their low confidence values in consensus trees. This is most likely to be an artifact of the jackknifing process since it not only disturbs gene adjacencies, but also causes the change of gene content.

Isolating different genes (but keeping them as separate chromosomes), in contrast, resamples the data in a way which better mimics evolution without gene content loss. We examined the trees obtained from the above experiments and calculated the number of low support branches (lower than 85%) inferred by jackknifing and the isolation method. Since the previous paper of Shi and Tang used EDE distances to obtain trees, to exclude the impact of the underlying phylogeny methods, we computed trees using both EDE and CDCJ distances.

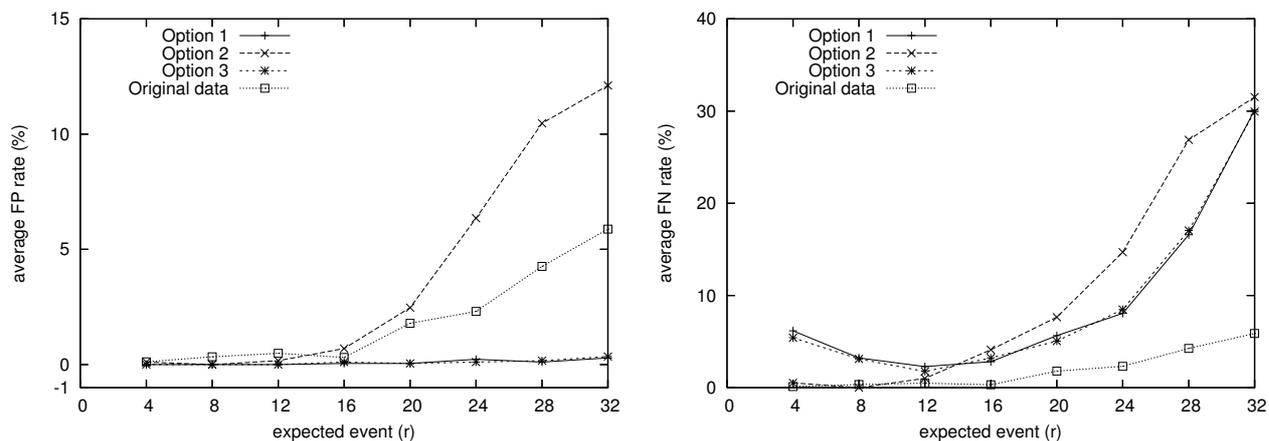


Fig. 4. (Top) FP of three options of isolating. (Bottom) FN of three options of isolating

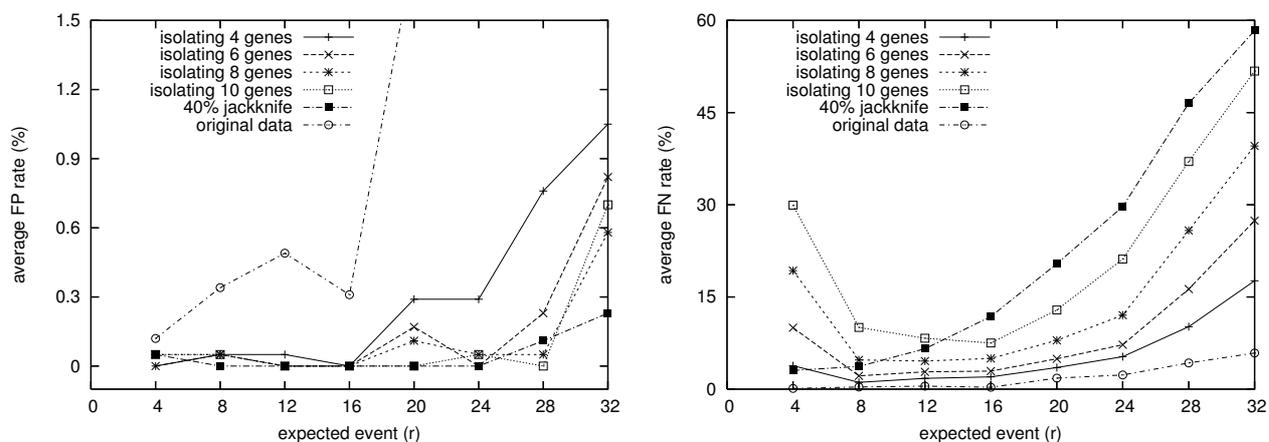


Fig. 5. Results of isolating various number of genes, using datasets with 20 genomes and 100 genes. (Top) FP rates. (Bottom) FN rates.

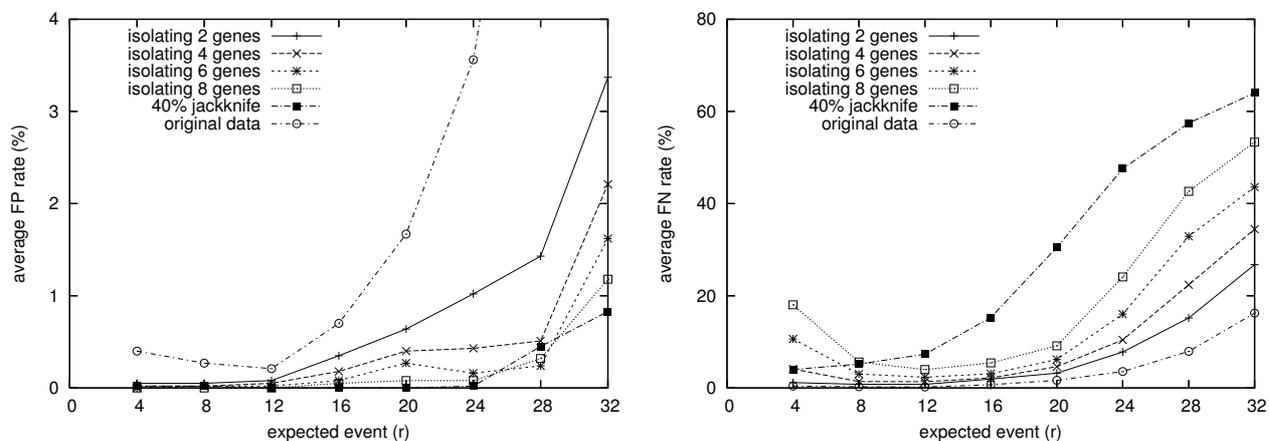


Fig. 6. Results of isolating various number of genes, using datasets with 40 genomes and 100 genes. (Top) FP rates. (Bottom) FN rates.

Fig 7 shows that our proposed method produces fewer low support branches than jackknifing does. Even for the most difficult datasets ($r = 32$), only $\sim 20\%$ branches have low confidence values. Combined with the low FP rates shown in Fig. 5 and 6, it is still safe to conclude that branches with

higher than 85% confidence values can be trusted. On the other hand, more good branches being preserved demonstrates that better consensus trees can be obtained by using the isolating method.

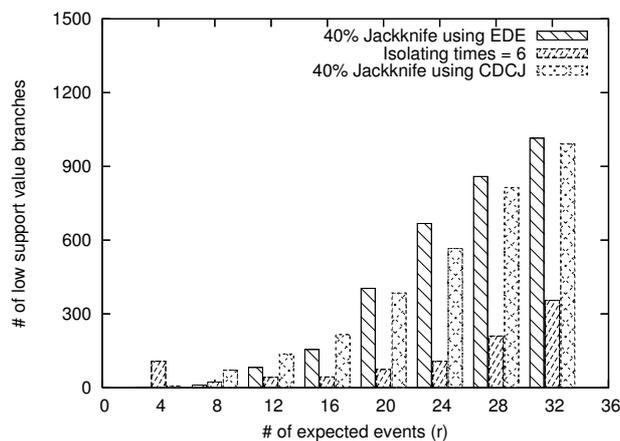


Fig. 7. The comparison of the number of low support branches inferred by jackknifing and isolation using 85% threshold, on datasets with 20 genomes. For each value of r , there are 100 datasets, thus 1700 internal branches

V. CONCLUSIONS

In this paper, we proposed a new resampling method based on a procedure we call gene isolation and presented a simple formula to determine the appropriate amount of isolation to use. Our simulation results show that this new method outperforms jackknifing by reducing the removal of valid branches while having a similarly low occurrence of confirming an incorrect branch.

VI. ACKNOWLEDGMENTS

The authors were supported by US National Institutes of Health (grant number 5R01GM078991-03) and National Science Foundation (grant number OCI 0904179). 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. and P. Holland (2000). Rare genomic changes as a tool for phylogenetics. *Trends in Ecology and Evolution*, 15, 454–459.
- [2] Felsenstein, J. (1985) Confidence limits on phylogenies: An approach using the bootstrap. *Evolution* 39, 783–791.
- [3] Moret, B. and T. Warnow (2005). Advances in phylogeny reconstruction from gene order and content data. *Methods in Enzymology* 395, 673–700.
- [4] Belda, E., A. Moya and F. Silva (2005). Genome rearrangement distances and gene order phylogeny in γ -Proteobacteria. *Mol. Biol. Evol.*, 22, 1456–1467.
- [5] Luo, H., J. Shi, W. Arndt, J. Tang and R. Friedman (2008). Gene order phylogeny of the genus *Prochlorococcus*. *PLoS ONE* 3, e3837.
- [6] Shi, J., Y. Zhang, H. Luo and J. Tang (2010). Using jackknife to assess the quality of gene order phylogenies. *BMC Bioinformatics*, 11, 168.
- [7] Yancopoulos, S., O. Attie and R. Friedberg, (2005). Efficient sorting of genomic permutations by translocation, inversion and block interchange. *Bioinformatics*, 21 (16), 3340–3346.
- [8] Hannenhalli, S. and P.A. Pevzner (1995). Transforming cabbage into turnip (polynomial algorithm for sorting signed permutations by reversals). *Proc. 27th Ann. Symp. Theory of Computing STOC'95*, 178–189.
- [9] Wang, L., R. Jansen, B. Moret, L. Raubeson and T. Warnow (2006). Distance-based genome rearrangement phylogeny. *J. Mol. Evol.* 63, 473–483.
- [10] Lin, Y. and B. Moret (2008). Estimating true evolutionary distances under the DCJ model. *Bioinformatics*, 24, i114–i122.

- [11] Saitou, N. and M. Nei (1987). The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.*, 4, 406–425.
- [12] Desper, R. and O. Gascuel (2002). Fast and accurate phylogeny reconstruction algorithms based on the minimum evolution principle. *J. Comput. Biol.*, 9, 687–705.
- [13] Larget, B., J. Kadane and D. Simon (2005). A Bayesian approach to the estimation of ancestral genome arrangements. *Mol. Phy. Evol.* 36, 214–223.
- [14] Moret, B., S. Wyman, D. Bader, T. Warnow and M. Yan (2001). A new implementation and detailed study of breakpoint analysis. *Proceedings of the 6th Pacific Symp. on Biocomputing (PSB'01)*, 583–594.
- [15] Bourque, G. and P. Pevzner (2002). Genome-scale evolution: reconstructing gene orders in the ancestral species. *Genome Research*, 12, 26–36.
- [16] Robinson, D. and L. Foulds (1981). Comparison of phylogenetic trees. *Mathematical Biosciences*, 53, 131–147.