

Selecting Genomes for Reconstruction of Ancestral Genomes

Guoliang Li¹, Jian Ma², and Louxin Zhang³

¹ Department of Computer Science
National University of Singapore (NUS), Singapore 117543
ligl@comp.nus.edu.sg

² Center for Biomolecular Science and Engineering
University of California at Santa Cruz
Santa Cruz, USA
jianma@soe.ucsc.edu

³ Department of Mathematics, NUS, Singapore 117543
matzlx@nus.edu.sg

Abstract. It is often impossible to sequence all descendent genomes to reconstruct an ancestral genome. In addition, more genomes do not necessarily give a higher accuracy for the reconstruction of ancestral character states. These facts lead to studying the genome selection for reconstruction problem. In this work, two greedy algorithms for this problem are proposed and tested on computer simulation data as well as a biological example.

1 Introduction

With more and more genomes having been sequenced, reconstructing ancestral proteins and genomic sequences becomes a popular approach for understanding the molecular origins and evolution of key components of virus, bacteria and eukaryotic organisms. Ancestral protein sequences for ribonuclease [8,21], Tu elongation factors [7], and steroid receptors [17] have been reconstructed and validated experimentally. Partial or complete DNA sequences for the common ancestor of placental mammals [1,11], HIV [6], and the 1918 flu virus [15] have also been constructed.

Parsimony, maximum likelihood and Bayesian methods are used for the reconstruction of ancestral protein or DNA sequences (see [4] for details of these methods). The reconstruction accuracy of these methods has been assessed by both theoretical analysis [13,19,14] and random simulation [20,1,2,18]. These analyses indicate that the topology of the phylogenetic tree relating the extant genomes to the target ancestral genomes affects the reconstruction accuracy significantly. For example, a starlike phylogeny allows the ancestral character states to be more accurately inferred than other topologies [14,3] although the actual situation is much more complicated [10]. Intuitively, more genomes should give better reconstruction accuracy at the root of a phylogeny. However, this is not always true even for a simple method like parsimony. Recently, we have shown

that, in many phylogenetic trees, the accuracy of the ancestral state in the root reconstructed with all the genomes is smaller than the accuracy of the ancestral state reconstructed with only one genome (Refer to Section 3, and also see [9] for details). This motivates us to study the following computational problem:

Given a phylogeny P on a set of genomes, an integer k and a reconstruction method \mathcal{M} , find a subset of k genomes in the phylogeny that gives the highest accuracy of reconstructing the ancestral genome at the root of the phylogeny, using method \mathcal{M} .

Another motivation for studying this problem is that, due to resource constraint, it is often impossible to sequence all the extant genomes that are evolved from the target ancestral genome. In this paper, we study the above genome selection for reconstruction problem. We develop two greedy algorithms for it and test them with the Fitch method on random simulation data as well as a biological example.

The rest of this paper is divided into six sections. In Section 2, we briefly introduce the Fitch method and its accuracy analysis in a simple Jukes-Cantor model. In Section 3, we demonstrate that more genomes are not necessarily better in accuracy for reconstructing an ancestral genome. In Section 4, we present two greedy algorithms for the genome selection for reconstruction problem. In Section 5, we test our algorithms against random phylogenetic trees. In Section 6, we examine a biological example. In Section 7, we conclude the paper with a few of remarks.

2 Parsimony Methods and Its Accuracy

2.1 A Simple Jukes-Cantor Evolutionary Model

Given the phylogenetic tree for a group of species, we assume that the character evolves by a Markov process, starting with a state at the root and proceeding to the leaves node by node. The probability that a node x receives a state t_x depends only on its parent node p and the conditions along the branch from p to x . The evolutionary model specifies the probability that a character c evolves to a character d on a branch from p to x as a conditional probability $\Pr[s_x = d | s_p = c]$. Here, we consider a simple Jukes-Cantor model. In this symmetric model, there are only two states, say 0 and 1, and the probability of a substitution change of any sort on any branch would be the same.

2.2 Parsimony Reconstruction Method

For reconstructing character evolution, parsimony methods assign to each internal node those states that allow for the fewest number of substitutions throughout the tree. In this paper, we study the genome selection for reconstruction problem with respect to the parsimony method proposed by Fitch [5]. This parsimony method assigns a set of states to each node one by one downward through the tree, starting with the leaves and using the subsets previously computed for

the node's children. For each leaf node, the observed state forms the state set. Assume A is an internal node with children B and C . The following rule is used to compute the state subset S_A from the state subsets S_B and S_C :

$$S_A = \begin{cases} S_B \cup S_C & \text{if } S_B \cap S_C = \phi, \\ S_B \cap S_C & \text{if } S_B \cap S_C \neq \phi. \end{cases}$$

The state set at the root contains all the possible states that will be assigned to it. We say that the method unambiguously reconstructs a state at the root if the state set contains only that state and ambiguously reconstructs a state if the state set contains both 0 and 1.

Note that the method presented in [13] (see also [12]) reconstructs the states of the internal nodes based on the information from all the leaf nodes, which is a little bit more complicated than the method described above. As far as the accuracy of the root is concerned, it gives the same state set as the method described above and hence has the same reconstruction accuracy at the root.

2.3 Reconstruction Accuracy

Assume the character evolves in a phylogeny with the root A according to a probabilistic evolutionary model. The evolutionary model specifies a prior probability for each state at A . When we say D is a state configuration at the leaves, we mean that it contains a state for each leaf in the phylogenetic tree. For a state c and a state configuration D at the leaves, we let $P[D|c]$ to denote the probability that the state c at the root evolves into the states given by D at the leaves in the phylogeny. Then the reconstruction accuracy of a method M is

$$P_{accuracy} = \sum_{c,D} \text{prior}(c)P[D|c]I(c, D, M),$$

where $I(c, D, M) = 1$ if the method M reconstructs c correctly from D at the root and 0 otherwise.

In this paper, we consider a symmetric evolutionary mode with two states 0 and 1. Hence, the reconstruction accuracy is independent of the prior distribution of the states. The unambiguous reconstruction accuracy of the Fitch method is

$$P_{accuracy} = \sum_D P[D|0]I(0, D, M) = \sum_D P[D|1]I(1, D, M).$$

There are three different state subsets $\{0\}$, $\{1\}$ and $\{0, 1\}$ with two states 0 and 1. For a state set t , and a state s , we use $P_N[t|s]$ to denote the probability that the state set t is computed at the node N by the Fitch's method given the true state s at N . It is not hard to see that $P_{accuracy} = P_A[\{0\}|0] = P_A[\{1\}|1]$.

At a leaf x with observed state s , we have

$$P_x[\{s\}|s] = 1, \quad P_x[\{s'\}|s] = P_x[\{0, 1\}|s] = 0$$

for $s' \neq s$. Let N be an internal node with the children L and R . Then, for $c, d = 0, 1$,

$$\begin{aligned}
 &P_N[\{d\}|c] \\
 &= \sum_{x,y=0,1} \Pr[s_L = x|s_N = c] \Pr[s_R = y|s_N = c] P_L[\{d\}|x] P_R[\{d\}|y] \\
 &\quad + \sum_{x,y=0,1} \Pr[s_L = x|s_N = c] \Pr[s_R = y|s_N = c] \\
 &\quad \times \{P_L[\{d\}|x] P_R[\{0, 1\}|y] + P_L[\{0, 1\}|x] P_R[\{d\}|y]\}
 \end{aligned}$$

and

$$P_N[\{0, 1\}|c] = 1 - P_N[\{0\}|c] - P_N[\{1\}|c].$$

The above recurrence relations give immediately a dynamic programming approach for computing the reconstruction accuracy of the Fitch’s method, which is used in our analysis in the rest of this paper. Such a method first appeared in [13].

The ambiguous reconstruction accuracy of the method takes the ambiguous state into consideration and is defined as

$$P_{A\text{-accuracy}} = P_A[\{1\}|1] + \frac{1}{2} P_A[\{0, 1\}|1]$$

where the first term is the unambiguous reconstruction accuracy and the second term in the expression simply says that, when either state 0 or 1 is equally parsimonious as a root state, we select either state with equal probability.

3 More Genomes Are Not Necessarily Better

Counterintuitively, more genomes do not necessarily give better reconstruction even for the parsimony methods [9]. The reason is that the reconstruction accuracy is highly sensitive to the topology used for the reconstruction and more genomes may introduce more noise in the reconstructed ancestral state. For completeness, we briefly summarize the partial results proved in [9] in this section.

We first consider the complete phylogenetic trees. Let T be the complete phylogeny with 4 leaves shown in Figure 1(a). We assume the conservation probability is p on any branch in T and $q = 1 - p$. For each node N , we denote the true state at N by s_N . Then, the conservation probability on each path from the root to a leaf is

$$\begin{aligned}
 &P_{path} \\
 &= \Pr[s_x = 1|s_A = 1] \\
 &= \Pr[s_B = 1|s_A = 1] \Pr[s_x = 1|s_B = 1] + \Pr[s_B = 0|s_A = 1] \\
 &\quad \times \Pr[s_x = 1|s_B = 0] \\
 &= p^2 + q^2,
 \end{aligned}$$

where we assume x is a leaf below the node B .

Let t_N be the reconstructed state set at a node N . For $V = B, C$ and $s, s' = 0, 1$,

$$\Pr[t_V|V = s] = \begin{cases} q^2 & \text{if } t_V = \{s'\} \text{ and } s' \neq s, \\ p^2 & \text{if } t_V = \{s'\} \text{ and } s' = s, \\ 2pq & \text{if } t_V = \{0, 1\}. \end{cases}$$

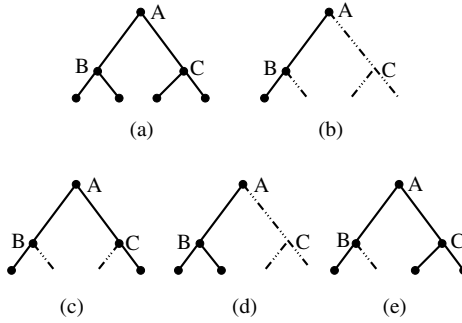


Fig. 1. (a) The complete phylogeny with 4 leaves; (b) The path topology; (c) The Λ -shape topology; (d) The Y-shape topology; (e) The W-shape topology

By this formula, the unambiguous reconstruction accuracy of using all the four taxa is

$$\begin{aligned}
 &P_{whole} \\
 &= \Pr[t_A = \{1\} | s_A = 1] \\
 &= \sum_{x,y \in \{0,1\}} \Pr[s_B = x | s_A = 1] \Pr[s_C = y | s_A = 1] \Pr[t_B = \{1\} | s_B = x] \\
 &\quad \times \Pr[t_C = \{1\} | s_C = y] \\
 &\quad + \sum_{x,y \in \{0,1\}} \Pr[s_B = x | s_A = 1] \Pr[s_C = y | s_A = 1] \\
 &\quad \times \{\Pr[t_B = \{1\} | s_B = x] \Pr[t_C = \{0, 1\} | s_C = y] \\
 &\quad + \Pr[t_B = \{0, 1\} | s_B = x] \Pr[t_C = \{1\} | s_C = y]\} \\
 &= (p^6 + q^6 + 2p^3q^3) + 2[2p^5q + 2pq^5 + 2p^2q^4 + 2p^4q^2] \\
 &= (p^3 + q^3)^2 + 4[p^2q(p^3 + q^3) + pq^2(p^3 + q^3)] \\
 &= (p^2 + q^2 - pq)(1 + pq).
 \end{aligned}$$

Since

$$\begin{aligned}
 &P_{path} - P_{whole} \\
 &= (p^2 + q^2) - (p^2 + q^2 - pq)(1 + pq) \\
 &= -(p^2 + q^2)pq + pq(1 + pq) \\
 &= 3p^2q^2.
 \end{aligned}$$

we have $P_{path} > P_{whole}$ unless $p = 0, 1$. Similarly, we can also show that the unambiguous reconstruction accuracy of using the topologies in Figure 1(c), 1(d) and 1(e) is smaller than P_{path} .

To find out how often the accuracy of using the whole phylogenetic tree to reconstruct ancestral character states at the root is smaller than the conservation probability on a path from the root to a leaf, we conducted simulation test by generating random phylogenetic trees in the Yule model.

In the Yule model, a random phylogenetic tree grows recursively from a single root node. In each step, one leaf in the current tree is selected to add two children with equal probability. The procedure repeats until a phylogenetic tree with the required number of leaves is generated.

For each set (N, p) of parameters, we generated five thousand random phylogenetic trees and count how many trees have the reconstruction accuracy less

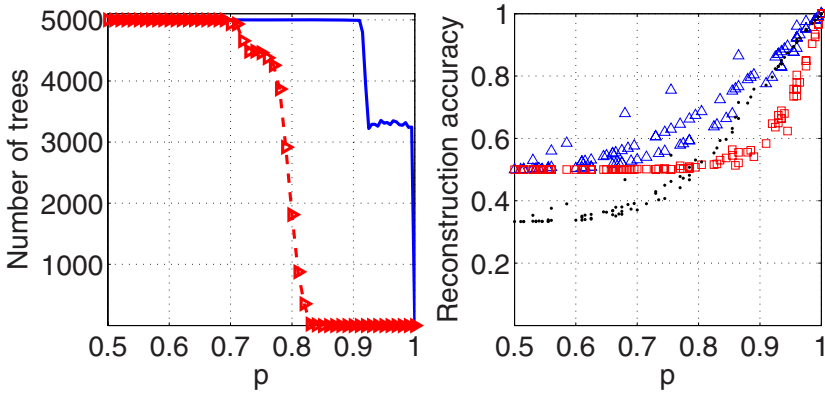


Fig. 2. (a) The number of the random phylogenetic trees in which the unambiguous reconstruction accuracy from the longest (triangle line) or shortest (dot line) path is better than the unambiguous reconstruction accuracy from the whole phylogeny. (b) The unambiguous reconstruction accuracy of the whole tree (dot curve), shortest path (triangle curve) and longest path (square curve).

than the conservation probability on the shortest or longest path from the root to a leaf. Here, N denotes the number of leaves in the random trees and is set to nine, fifteen, or twenty; p represents the conservation probability on each branch and is set to $0.5 + 0.01i$ for each $0 \leq i \leq 49$.

Figure 2(a) shows the number of randomly-generated phylogenetic trees in which the conservation probability on the shortest path or longest path is larger than the reconstruction accuracy of using the whole tree. When the conservation probability p on each branch is in the range of 0.5 and 0.8, the conservation probability on the shortest path to some leaf is better in most of the trees. When p exceeds 0.9, the number of ‘bad’ trees decreases rapidly. Figure 2(b) shows sampled reconstruction accuracy of these three different reconstructions.

We show the fact the more genomes do not always give better unambiguous reconstruction accuracy in some phylogenetic trees. This observed fact also holds for the ambiguous reconstruction accuracy [9].

4 Algorithms for Genome Selection

The counterintuitive observation in above section and the fact that limited resource prohibits one to sequence all the descendent genomes for ancestral reconstruction motivate us to study the genome selection for reconstruction problem. Formally, this problem is defined as

Genome selection for reconstruction

Instance: A phylogenetic tree P on a set of n genomes, a number k and a reconstruction method \mathcal{M} .

Question: Find k genomes in P that allows the ancestral character states at the root of P to be reconstructed with the maximum accuracy, using method \mathcal{M} .

Since the reconstruction accuracy depends on both the topology of the given phylogeny and the conservation probability of each branch, the genome selection for reconstruction problem is unlikely polynomial-time solvable although its NP-hardness is not proved yet. In the rest of this section, we present two greedy algorithms for it.

4.1 Forward Greedy Algorithm

The forward greedy algorithm selects the k genomes one by one based on accuracy increment. Initially, the algorithm chooses the genome that has the shortest evolutionary distance from the root. In each of the following $k - 1$ steps, the algorithm selects a genome that gives the maximum increment on reconstruction accuracy. In summary, the forward greedy algorithm can be described as follows:

FORWARD GREEDY ALGORITHM

1. Set $S \leftarrow \phi$;
2. Add the nearest genome to S ;
3. For $i = 1, 2, \dots, k - 1$ do {
 - for each genome g not in S , compute the accuracy A_g of the reconstruction by applying \mathcal{M} to $S \cup \{g\}$;
 - Add g to S if A_g is the maximum over all g s;
4. Output S .

4.2 Backward Greedy Algorithm

The backward greedy algorithm removes $n - k$ genomes one by one by considering the accuracy decrease. Initially, there are n genomes. In each of $n - k$ steps, the algorithm selects a genome whose removal leads to the least decrease in reconstruction accuracy.

BACKWARD GREEDY ALGORITHM

1. Let S contain all the genomes in the phylogeny;
2. For $i = 1, 2, \dots, n - k$ do {
 - for each genome g in S , compute the accuracy A_g of the reconstruction by applying \mathcal{M} to $S - \{g\}$;
 - Remove g from S if A_g is the maximum over all g 's;
3. Output S .

Since the backward greedy algorithm starts from the full phylogeny, it is not hard to see that the backward greedy algorithm is not efficient as the forward greedy

algorithm especially when reconstruction method such as the maximum likelihood, is used. However, as we will see below, this method has better performance.

5 Simulation Test

To evaluate the performance of the forward and backward greedy algorithms, we apply them with the Fitch method on random phylogenetic trees generated in the Yule model. For $p = 0.75, 0.80, 0.85, 0.90, 0.95, 0.99$ and $N = 9, 16$, we respectively generated one hundred balanced and one hundred imbalanced random trees with N leaves using the method described in the previous section.

For each random tree with nine leaves, we apply the two greedy algorithms to find a three-leaf subset and a six-leaf subset; for each random tree with sixteen leaves, we apply the two greedy algorithms to find a five-leaf subset and a ten-leaf subset. The accuracy of reconstructing the character states at the root using the found subset is computed and compared with the optimal accuracy over all the subsets containing the desired number of genomes and the accuracy of using all the genomes. Figure 3 shows the average accuracies from different algorithms on one hundred balanced random trees. The left bar graph is the average accuracy of six-leaf subsets from the balanced random phylogeny with nine leaf nodes, and the right bar graph is the average accuracy of ten-leaf subsets from the balanced random phylogeny with sixteen leaf nodes. The performance of the greedy algorithms on the imbalanced trees with $p < 0.9$ is generally better (data not shown here due to space limitation), which is consistent with the results in Section 3.

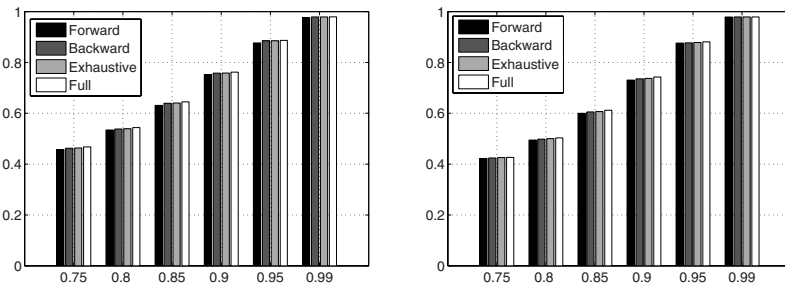


Fig. 3. The left and right bar graphs summarize the average reconstruction accuracy of the subsets found by the two algorithms against the optimal accuracy on the randomly generated phylogenetic trees with nine and sixteen leaves respectively

In these tests, the algorithms can identify subsets of genomes that result in better reconstruction accuracy than obtained using all genomes in the tree. When the conservation probability $p = 0.75$, the accuracy from the backward greedy algorithm on the three-leaf subset of the nine-leaf phylogeny is always better than the accuracy from the full phylogeny (data not shown). As the conservation probability increases, the greedy algorithms obtain better accuracy less frequently.

Both tests also indicate that the backward greedy algorithm yields higher reconstruction accuracy than the forward greedy algorithm in about 80% random trees. But, as we mentioned earlier, the drawback of the backward greedy algorithm is that it is time-consuming, especially when the phylogenetic tree is large and maximum likelihood or a Bayesian method is used for reconstruction.

Furthermore, Figure 3 shows that, on average, the accuracy from the greedy algorithms are comparable to, if not better than, the accuracy from the full phylogeny. This provides the support for selecting a subset of the genomes to reconstruct the ancestral genomes when there are resource constraints and we can not sequence all the extant genomes in the domain of interest.

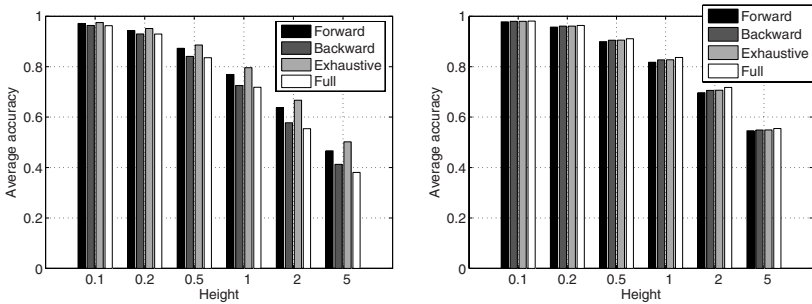


Fig. 4. Average accuracy under different tree heights. The left and right bar graphs summarize the average unambiguous reconstruction accuracy and ambiguous reconstruction of the subsets found by the two algorithms against the optimal accuracy from exhaustive search on the randomly generated phylogenetic trees with three-leaf subsets on nine-leaf phylogeny respectively. The x-axis is the height of the trees and the y-axis is the average accuracy under the individual heights.

We also generated random trees with different heights using Evolver in the PAML package (<http://abacus.gene.ucl.ac.uk/software/paml.html>). We considered the trees with nine and sixteen leaves. The parameters used to generate the trees are: 10 for Birth rate, 5 for Death rate, 1 for Sampling fraction, and 0.1, 0.2, 0.5, 1, 2, 5 for height. The height means the sum of the branch lengths from the root to all leaf nodes. For each possible combination of parameter values, we generated one hundred random trees and estimated the transition probability along each branch using the Jukes-Cantor model. The left panel of Figure 4 shows the average unambiguous reconstruction accuracy for different heights. The right panel of Figure 4 shows the average ambiguous reconstruction accuracy for different heights. In both cases, the solutions output by our greedy algorithms are near optimal. Note also that, for ambiguous reconstruction accuracy, the backward greedy algorithm outperforms the forward greedy algorithm.

Unlike the unambiguous reconstruction accuracy, the ambiguous reconstruction accuracy from the full phylogeny is better on average, indicating that more genomes introduce more noise. It seems true that more genomes always result in higher ambiguous reconstruction accuracy in an ultrametric phylogenetic tree

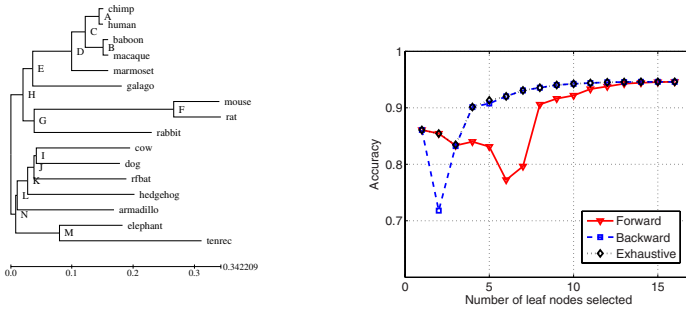


Fig. 5. A phylogenetic tree in the reconstruction of the Boreoeutherian ancestor and the unambiguous reconstruction accuracies of the greedy algorithms on this tree. In the left graph, the branch lengths are the substitution rate. In the right graph, the solid line with triangles represents the accuracies from the forward greedy algorithm, the dashed line with squares represents the accuracies from the backward greedy algorithm, and the dotted line with diamonds represents the accuracies from the exhaustive search.

in which all the paths to a leaf have the equal height. However, it is not known how to prove this hypothesis.

6 A Biological Example

We consider the reconstruction of the so-called Boreoeutherian ancestor where a rapid radiation of many different lineages occurred. We applied the both forward and backward algorithms to the phylogeny shown in Figure 5 (see [16]). Among the sixteen extant species of this phylogenetic tree, the genomes of human, chimp, macaque, rat, mouse, and dog have been sequenced for the first time; other genomes have been partially sequenced.

In this example, we examined the expected accuracy of the reconstruction of the four nucleotides on each base in the Boreoeutherian ancestor. The branch weight in the phylogeny is the substitution rate. Therefore, under the Jukes-Cantor model, we assume that, for each branch, the conservation probability is one minus the branch weight and the probability of one nucleotide replacing another is one third of the substitution rate. Since the true ancestral nucleotide residues are unknown at the Boreoeutherian ancestor, it is impossible to obtain the true reconstruction accuracy. As a result, we calculated the expected reconstruction accuracy using the formula stated in Section 2.3. (Here, we considered four states, rather than two states in the model used in section 2 and section 3.)

For each of k from one to sixteen, the reconstruction accuracy obtained using k genomes, estimated by the greedy algorithms, are compared in Figure 5. When $k = 1, 2$, the forward greedy algorithm performed similarly to the exhaustive search algorithm. When $k = 3$, all three algorithms obtained similar accuracy. When $k > 3$, the performance of the backward greedy algorithm is similar to the exhaustive search algorithm, and the performance of the forward greedy

algorithm is worse. For example, when $k = 8$, the backward algorithm output the following genomes: human, dog, galago, mouse, rabbit, dog, armadillo, elephant, leading to the unambiguous reconstruction accuracy as high as 93.6%, which is quite near the accuracy 94.6% obtained using the full phylogeny.

7 Conclusion

It is well known that parsimony method is not consistent when the branches are long more characters do not lead to the right phylogeny (see Chapter 9 of [4] for details). Here, we observe that more genomes are not necessarily better in the reconstruction of ancestral character states with a given phylogeny, giving a complementary example in which more data is not necessarily better.

Motivated by the above counterintuitive result and the impossibility of sequencing all the descendent genomes for ancestral genome reconstruction, we have studied the genome selection for reconstruction problem in this work. We proposed two greedy algorithms for the problem and tested them with simulation data. The experiment results showed that, in most of the cases, the accuracy from the greedy algorithms is comparable to the highest accuracy of using the same number of genomes; it is also comparable to, if not better than, the accuracy of using all the genomes in the full phylogeny. In general, the forward algorithm is more straightforward, but has poor performance compared with the backward greedy algorithm.

We also tested our algorithms on the reconstruction of the Boreoeutherian ancestor of the placental mammals. The test shows that using only eight genomes identified by the backward greedy algorithm, an expected reconstruction accuracy of 93.6% can be obtained. It is quite close to the accuracy obtained with the full phylogeny, namely 94.6%. This indicates that selecting the genomes for ancestral genome reconstruction is also practical.

Acknowledgment

The authors would like to thank the reviewers and D. Durand for their valuable suggestions on revising the paper. LX Zhang gratefully acknowledged the NUS ARF grant R-146-000-068-112 and NSFChina3052802 for partially supporting this project. He also thanks Webb Miller for stimulating this research by pointing out the paper [10] to him.

References

1. Blanchette, M., Green, E.D., Miller, W., Haussler, D.: Reconstructing large regions of an ancestral mammalian genome in silico. *Genome Res.* 14, 2412–2423 (2004)
2. Cai, W., Pei, J.M., Grishin, N.V.: Reconstruction of ancestral protein sequences and its application. *BMC Evol. Biol.* 4, e33 (2004)
3. Evens, W., Kenyon, C., Peres, Y., Schulman, L.J.: Broadcasting on trees and the ising model. *Annals of Applied Prob.* 10, 410–433 (2000)

4. Felsenstein, J.: *Inferring Phylogenies*, Sinauer Associates. Sunderland, Massachusetts (2004)
5. Fitch, W.M.: Toward Defining the Course of Evolution: Minimum Change for a Specific Tree Topology. *Systematic Zoology* 20, 406–416 (1971)
6. Hillis, D.M., Huelsenbeck, J.P., Cunningham, C.W.: Application and accuracy of molecular phylogenies. *Science* 264, 671–677 (1994)
7. Gaucher, E.A., Thomson, J.M., Burgan, M.F., Benner, S.A.: Inferring the palaeoenvironment of ancient bacteria on the basis of resurrected proteins. *Nature*, 285–288 (2003)
8. Jermann, T.M., Opitz, J.G., Stackhouse, J., Benner, S.A.: Reconstructing the evolutionary history of the artiodactyl ribonuclease superfamily. *Nature* 374, 57–59 (1995)
9. Li, G.L., Steel, M., Zhang, L.X.: More taxa are not necessarily better for the reconstruction of ancestral sequence by parsimony. Manuscript (2007)
10. Lucena, B., Haussler, D.: Counterexample to a claim about the reconstruction of ancestral character states. *Syst. Biol.* 54, 693–695 (2005)
11. Ma, J., Zhang, L., Suh, B.B., Raney, B.J., Burhans, R.C., Kent, W.J., Blanchette, M., Haussler, D., Miller, W.: Reconstructing contiguous regions of an ancestral genome. *Genome Res.* 16, 1557–1565 (2006)
12. Maddison, W.P., Maddison, D.R.: *MacClade: analysis of phylogeny and character evolution*. Version 3, Sinauer, Sunderland, MA
13. Maddison, W.P.: Calculating the probability distributions of ancestral states reconstructed by parsimony on phylogenetic trees. *Systematic Biology* 44, 474–481 (1995)
14. Schultz, T.R., Cocroft, R.B., Churchill, G.A.: The reconstruction of ancestral character states. *Evolution* 50, 504–511 (1996)
15. Taubenberger, J.K., Reid, A.H., Lourens, R.M., Wang, R., Jin, G., Fanning, T.: Characterization of the 1918 influenza virus polymerase genes. *Nature* 437, 889–893 (2005)
16. The ENCODE Project Consortium, Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project, *Nature*, 447, 799–816 (2007)
17. Thornton, J.W., Need, E., Crews, D.: Resurrecting the ancestral steroid receptor: ancient origin of estrogen signaling. *Science* 301, 1714–1717 (2003)
18. Williams, P.D., Pollock, D.D., Blackburne, B.P., Goldstein, R.A.: Assessing the accuracy of ancestral protein reconstruction methods. *PLoS Comput Biol.* 2, e69 (2006)
19. Yang, Z.H., Kumar, S., Nei, M.: A new method of inference of ancestral nucleotide and amino acid sequences. *Genetics* 141, 1641–1650 (1995)
20. Zhang, J., Nei, M.: Accuracies of ancestral amino acid sequences inferred by parsimony, likelihood, and distance methods. *J. Mol. Evol.* 44(S1), 139–146 (1997)
21. Zhang, J., Rosenberg, H.F.: Complementary advantageous substitutions in the evolution of an antiviral RNase of higher primates. *Proc. Natl. Acad. Sci. USA* 99, 5486–5491 (2002)