**BMC
Bioinformatics**

RESEARCH ARTICLE

Open Access

Analysis on the reconstruction accuracy of the Fitch method for inferring ancestral states

Jialiang Yang, Jun Li, Liuhuan Dong, Stefan Grünewald[*]

Abstract

Background: As one of the most widely used parsimony methods for ancestral reconstruction, the Fitch method minimizes the total number of hypothetical substitutions along all branches of a tree to explain the evolution of a character. Due to the extensive usage of this method, it has become a scientific endeavor in recent years to study the reconstruction accuracies of the Fitch method. However, most studies are restricted to 2-state evolutionary models and a study for higher-state models is needed since DNA sequences take the format of 4-state series and protein sequences even have 20 states.

Results: In this paper, the ambiguous and unambiguous reconstruction accuracy of the Fitch method are studied for N-state evolutionary models. Given an arbitrary phylogenetic tree, a recurrence system is first presented to calculate iteratively the two accuracies. As complete binary tree and comb-shaped tree are the two extremal evolutionary tree topologies according to balance, we focus on the reconstruction accuracies on these two topologies and analyze their asymptotic properties. Then, 1000 Yule trees with 1024 leaves are generated and analyzed to simulate real evolutionary scenarios. It is known that more taxa not necessarily increase the reconstruction accuracies under 2-state models. The result under N-state models is also tested.

Conclusions: In a large tree with many leaves, the reconstruction accuracies of using all taxa are sometimes less than those of using a leaf subset under N-state models. For complete binary trees, there always exists an equilibrium interval [*a, b*] of conservation probability, in which the limiting ambiguous reconstruction accuracy equals to the probability of randomly picking a state. The value *b* decreases with the increase of the number of states, and it seems to converge. When the conservation probability is greater than *b*, the reconstruction accuracies of the Fitch method increase rapidly. The reconstruction accuracies on 1000 simulated Yule trees also exhibit similar behaviors. For comb-shaped trees, the limiting reconstruction accuracies of using all taxa are always less than or equal to those of using the nearest root-to-leaf path when the conservation probability is not less than $\frac{1}{N}$. As a result, more taxa are suggested for ancestral reconstruction when the tree topology is balanced and the sequences are highly similar, and a few taxa close to the root are recommended otherwise.

Background

Ancestral state reconstruction attempts to predict properties of ancestral proteins, genes and even whole genomes in a given phylogeny according to data of extant species. This approach to understanding protein functions and evolution was first proposed by Pauling and Zukerkandl in their seminal work [1]. Thereafter, with the increasing availability of biological data it has become a technique of growing importance in investigating the functions and origins of genes and proteins [2-9].

Parsimony and maximum likelihood (ML) are the two most popular criteria utilized to reconstruct ancestral states when the phylogenetic tree representing the evolutionary history of a character is known [6,10]. Parsimony methods minimize the total number of hypothetical substitutions along all branches of the evolutionary tree. The Fitch method was the first parsimony method for inferring ancestral states [11]. It is a linear time algorithm and is accurate for taxa with highly similar sequences. The method was later modified by Sankoff to account for different rates of substitutions among

* Correspondence: stefan@picb.ac.cn
CAS-MPG Partner Institute for Computational Biology, Key Laboratory of
Computational Biology, Shanghai Institutes for Biological Sciences, Chinese
Academy of Sciences, Shanghai 200031, PR China

states [12,13]. The reader is referred to a survey book [14] for reviews of parsimony methods and their variants. In contrast to parsimony methods, ML methods choose a state to be the ancestral state such that the observed states could have been evolved from it with maximum likelihood. ML inference of ancestral sequences was pioneered by Yang, Kumar and Nei [15] and by Koshi and Goldstein [16]. Later, a widely used variant of ML method called the Bayesian approach was introduced by Huelsenbeck and his coworkers [17,18]. The reader is referred to [6] and [19] for reviews of ML methods and their variants.

Due to the extensive usage of ancestral reconstruction methods, it has become a significant scientific endeavor to study their reconstruction accuracies. These accuracies of different methods have been either estimated by statistical simulations [20,21] or calculated precisely by theoretical analyses [22-27]. For example, under a 2-state Jukes-Cantor model, several recurrence systems for calculating the reconstruction accuracies of the Fitch algorithm were presented for a given phylogenetic tree [22,24,25,27]. It was shown in these studies that the reconstruction accuracies depend largely on the topology of the phylogenetic tree. Thus, reconstruction accuracies and their asymptotic properties on the number of leaves were also analyzed for extremal trees like complete binary trees and comb-shaped trees (or rooted caterpillars) [25,27]. However, by far most theoretical analyses have been limited to 2-state models. More effort should be made to study the reconstruction accuracies under higher-state models as there are 4 states for DNA sequences and even 20 states for protein sequences.

In this paper, we study the ambiguous and unambiguous reconstruction accuracy of the Fitch algorithm for reconstructing the root state under N-state evolutionary models. We first present a general recurrence system for calculating the reconstruction accuracies on any given phylogenetic tree. We developed software that implements this system. As pointed out by Li et al. [24], more taxa are not necessarily better for the reconstruction of ancestral states. Our recurrence system and software can be used to select good subsets of taxa to reconstruct ancestral DNAs, proteins, or other characters.

After that, we restrict the analyses to 3 extremal evolutionary trees under the N-state Jukes-Cantor model, namely equal-branch complete binary tree, equal-branch comb-shaped tree and Hennigain comb-shaped tree [24,28]. It is clear that for the equal-branch trees the substitution probability along any branch is the same and is denoted by p, thus the conservation probability is $q := 1 - (N - 1)p$. As examples, we analyze reconstruction accuracies and their asymptotic properties on the number of leaves for $N = 2, 4, 5, 20$. We also compare the limiting ambiguous and unambiguous reconstruction accuracy of using all taxa with those of using a nearest root-to-leaf path. Finally, 1000 Yule trees with 1024 leaves are generated by the software Mesquite [29] and analyzed to simulate real phylogenetic trees.

From the studies, we observe several interesting properties for the reconstruction accuracies under N-state models. First, for equal-branch complete binary trees, there always exists an equilibrium interval $[a, b]$ of conservation probability q such that, with the number of leaves tending to infinity, the ambiguous reconstruction accuracy converges to $\frac{1}{N}$, the reconstruction accuracy of randomly picking the ancestral state from N possible states. For example, the equilibrium interval for complete binary trees is $\left[\frac{1}{8}, \frac{7}{8}\right]$ under the 2-state Jukes-Cantor model [22,25,27]. However, a becomes 0 when $N \geq 3$. We calculate b for $N = 2, ..., 25$ and find that b decreases slowly with the increase of N. The reconstruction accuracies for 1000 Yule trees exhibit similar behaviors. Second, for any equal-branch comb-shaped tree, the limiting reconstruction accuracies using the nearest root-to-leaf path are always greater than those using all taxa if the conservation probability is larger than $\frac{1}{N}$. Finally, for any Hennigian comb-shaped tree, the limiting ambiguous reconstruction accuracy is always equal to that of randomly picking a state, $\frac{1}{N}$, whereas the limiting unambiguous one is equal to

$$\frac{N^{N-2}}{\sum_{i=l}^{N} N^{N-i} \frac{(N-1)!}{(N-x)!}}.$$

Our results suggest that more taxa should be used for reconstructing ancestral states if the tree topology is balanced and the sequences are highly similar, whereas some taxa close to the root are recommended if the tree topology is very unbalanced. In addition, under evolutionary models with molecular clock, the reconstructed state by the Fitch algorithm is as bad as a state randomly picked when the conservation probability is low or the phylogenetic tree is very unbalanced. The suggestions are also partially applicable for ML methods as ML inference of the root state is the same as that of maximum parsimony estimation under simple models such as Jukes-Cantor models when the branch lengths of the phylogenetic tree are unknown [30].

Methods
The Fitch Method
Let C be a character with state set $\mathcal{S}$ and T be a rooted phylogenetic tree in which each leaf is assigned a state in $\mathcal{S}$. The Fitch method infers the states of internal nodes in a two-stage process, namely the "leaf-to-root stage"

and the "root-to-leaf stage". In the first stage, it assigns a set of states S_u to each node u of T as follows:

(1) If u is a leaf, S_u contains only the state of u,
(2) If u is an internal node having children v and w,

$$S_u = S_v * S_w = \begin{cases} S_v \cap S_w & \text{if } S_v \cap S_w \neq \varnothing, \\ S_v \cup S_w & \text{otherwise.} \end{cases}$$

The assignment starts from the leaves, and proceeds to the internal nodes until the root r is assigned a subset S_r. In the second stage, only one state s_u is chosen from the state subset S_u at each node u. This is realized by first picking randomly a state in S_r, and proceeding downwards as follows. Suppose v is a child of u, then

$$s_v = \begin{cases} s_u & \text{if } s_u \in S_v, \\ \text{any state in } S_v & \text{otherwise.} \end{cases}$$

As a result, any state in S_r is chosen as the root state with an equal probability $\frac{1}{|S_r|}$, where $|S_r|$ denotes the cardinality of S_r.

Reconstruction Accuracies of the Fitch Method

Unambiguous and ambiguous reconstruction accuracy are two important criteria to evaluate the quality of a reconstruction method. Let C be a character with state set $\mathcal{S} = \{s_1, s_2, \cdots, s_N\}$. For simplicity, we order the states and restate $\mathcal{S} = \{1, 2, \cdots, N\}$. In order to study the reconstruction accuracies of the Fitch method on a given phylogenetic tree T with root r, a Markov model is usually assumed to represent the true biological evolutionary process of C, which specifies:

(1) $\Pr[t_r = i]$, that is, the initial probability that the root state is i for $i = 1, 2, \ldots, N$, and
(2) $\Pr[t_v = j \mid t_u = i]$, that is, the transition probability that a state i evolves to j along the branch from node u to v for any $I, j = 1, 2, \ldots, N$ and branch uv.

In particular, $\Pr[t_v = i \mid t_u = i]$ is called the *conservation probability* of i along uv, and $\Pr[t_v = j \mid t_u = i]$ with $i \neq j$ is called the *substitution probability* from i to j along uv. Clearly, the probability of a state in each node is already determined by the model. For any state i and vertex u in T, we use $\Pr[t_u = i]$ to denote the probability that the state of u is i. We assume throughout this paper that the evolutionary model is *symmetric* on all states, that is $\Pr[t_v = j \mid t_u = i] = \Pr[t_v = i \mid t_u = j]$ and $\Pr[t_v = i \mid t_u = i] = \Pr[t_v = j \mid t_u = j]$ for any two states i and j, and any branch uv.

After the character evolves from the root r under the model, the leaves will receive one of many possible distributions of the states each with some probability. The Fitch method is then applied to these distributions to infer ancestral states. Let Ψ be the set containing all the possible distributions of leaf states. For any state $i \in \mathcal{S}$ and any distribution $\mathcal{D} \in \Psi$, let $\Pr[\mathcal{D} \mid i]$ be the probability that the leaf nodes receive the distribution $\mathcal{D}$, given that the state at the root is i. Let $B \subseteq \mathcal{S}$, we further let $C(B, \mathcal{D})$ be the probability that the Fitch method reconstructs set B at root from $\mathcal{D}$. Since both the evolutionary model and the Fitch method are symmetric on all states, the reconstruction accuracies are independent of the prior distribution of initial states. Thus, we choose 1 to be the root state and the *unambiguous reconstruction accuracy* (UA) is defined as

$$UA = \sum_{\mathcal{D} \in \Psi} \Pr[\mathcal{D} \mid 1] C(\{1\}, \mathcal{D}), \tag{1}$$

the probability that it outputs the true state 1. Similarly, the *ambiguous reconstruction accuracy* (AA) is defined as

$$AA = \sum_{1 \in B} \frac{1}{|B|} \sum_{\mathcal{D} \in \Psi} \Pr[\mathcal{D} \mid 1] C(B, \mathcal{D}). \tag{2}$$

That is, if the reconstructed set B contains 1, there is still a probability of $\frac{1}{|B|}$ to infer the true root state.

Recurrence Relations to Calculate Reconstruction Accuracies

Let Z be an internal node with two children X and Y. Since the evolutionary model is symmetric on all N states, the substitution probability between any two states is the same along a given branch. We use p_X and p_Y to denote the substitution probabilities along branches ZX and ZY, respectively. Clearly, the corresponding conservation probabilities on any state are $1 - (N - 1)p_X$ and $1 - (N - 1)p_Y$. In the following, we derive a recursive system involving $2N - 1$ recursive formulas to calculate the reconstruction accuracies of a parent node from those of its two children.

Before we present the system, it is worthy of mentioning that the original dynamic programming approach in Maddison [23] can be applied to calculate both accuracies for N-state models. However, it involves the calculation of $2^N - 1$ recursive formulas to calculate the reconstruction accuracies of a parent node from those of its two children, which is not efficient when both the evolutionary tree and the number of states N are large. Here, we take the advantage of the symmetries of both the Fitch method and the evolutionary models. More precisely, the reconstructed root states are categorized into $2N - 1$ classes. Let $\mathcal{B}_{2i-1} = \{B \subseteq \mathcal{S} : 1 \in B \text{ and } |B| = i\}$ for $1 \leq i \leq N$

and $\mathcal{B}_{2i} = \{B \subseteq \mathcal{S} : 1 \notin B$ and $|B| = i\}$ for $1 \leq i \leq N$ - 1. Then $\mathcal{B} = \{\mathcal{B}_1, \ldots, \mathcal{B}_{2N-1}\}$ is a partition of the set of all non-empty subsets of S. For any node u, define

$$A_i^u = \sum_{\mathcal{D} \in \Psi} \mathrm{Pr}_u[\mathcal{D} \mid 1] C_u(\mathcal{B}_i, \mathcal{D}),$$

where $\mathrm{Pr}_u[\mathcal{D} \mid 1]$ denotes the probability that the leaf configuration under u is $\mathcal{D}$ given that the true state at u is 1, and $C_u(\mathcal{B}_i, \mathcal{D})$ denotes the probability that the reconstructed set at u from $\mathcal{D}$ is $\mathcal{B}_i$. By this definition,

$$UA = A_1^r \text{ and } AA = \sum_{k=1}^{N} \binom{N-1}{k-1} \frac{1}{k} A_{2k-1}^r.$$

For any node u, let the reconstructed state set be Bu. Then for any $B \in \mathcal{B}_i$, $BZ = B$ if and only if: (1) $B_X \cap B_Y = B$, or (2) $B_X \cap B_Y = \varnothing$ and $B_X \cup B_Y = B$. Thus A_i^Z can be calculated from the reconstruction accuracies on node X and Y in conjunction with the substitution or conservation probabilities along the two branches ZX and ZY (see Additional File 1 for details). Recurrence formulas for 2-state models can be found in [22-27]. We present the recurrence system and initial conditions for N-state models in Additional File 1. To facilitate our study, we also implemented a computer program which takes a phylogenetic tree in Newick format and the substitution rate along branches as inputs. The phylogenetic tree can be inferred by methods like Neighbor-Joining [31] and the substitution rate can also be estimated, see for example [32]. A potential application of our algorithm and program is to select the subsets of taxa to accurately reconstruct ancestral sequences such as DNAs and proteins.

Results and Discussion

As the reconstruction accuracies of the Fitch algorithm depend largely on the topology of phylogenetic trees, we focused our attention on two extremal tree shapes: complete binary trees which are the most balanced trees and comb-shaped trees (caterpillars), the most unbalanced trees. We are interested in trees with many taxa and in our results and figures we choose sufficiently many taxa to exhibit the asymptotic behavior. In order to simulate more realistic evolutionary scenarios, we also generated and analyzed 1000 Yule trees with 1024 leaves.

Reconstruction Accuracies for Equal-branch Complete Binary Trees

Let T_n be the equal-branch complete binary tree of 2^n leaves in which the substitution probability is p on each branch, and thus the conservation probability is $q = 1 - (N - 1)p$. Since the subtree rooted at each child of the root r is the complete binary tree of 2^{n-1} leaves, the recurrence system to calculate the reconstruction

accuracy can be simplified (see Additional File 2). We simulated the obtained system by Matlab and studied the asymptotic properties of ambiguous and unambiguous reconstruction accuracy by using all taxa as well as by using a root-to-leaf path. The results for $N = 2, 4, 5$ and 20 are plotted in Figure 1.

Similar to the 2-state case in [24], we observed that for $N = 4, 5$ and 20: (1) UA or AA on a root-to-leaf path is always less than or equal to AA on all the terminal taxa, and (2) UA or AA on a root-to-leaf path is greater than UA on all terminal taxa when q is small but becomes smaller than that when q is larger than a threshold. We conjecture that the two properties hold for arbitrary number of states.

It is shown in [25,27] that for 2-state models, UA and AA diverge when $q < \frac{1}{8}$ and converge to $\frac{1}{3}$ and $\frac{1}{2}$, respectively, when $\frac{1}{8} \leq q \leq \frac{7}{8}$, as n tends towards infinity. Surprisingly, we observed that the divergent interval vanishes for more than two states. In addition, in an interval $[0, b]$ AA converges to $\frac{1}{N}$, the probability of randomly picking a state. A possible reason is that in the 2-state scenario, the node state changes alternatingly with the increase of the level in complete binary trees. As a result, both UA and AA are alternating and thus divergent. However, their even and odd series converge to different limits (see [25,27]). For more-state models, the alternating property vanishes since there are more states to travel. If q is not very large, the probability for a node state to be one specific state becomes randomized as the level increases. Consequently, both UA and AA are convergent and AA converges to $\frac{1}{N}$. Clearly, b is a very important parameter. We listed in Table 1 the estimated values of b for $N = 2, \ldots, 25$. Table 1 shows that b decreases with the increase of N and it seems to converge to a number between 0.7 and 0.76. We offer the following explanation why more states make it easier for the Fitch algorithm to be better than random: with increasing number of states and constant conservation probability it is less likely that there are independent mutations from the root state to the same non-root state. Instead, it will happen more often that there are independent mutations from the root state to different non-root states. The latter situation makes it easier for the Fitch algorithm to reconstruct the ancestral state correctly. For example, if the reconstructed state sets at the children of an interior vertex Z are $\{1, i\}$ and $\{1, j\}$ (with $I, j \neq 1$), respectively, then the set reconstructed at Z is $\{1, i\}$ if $i = j$ and just $\{1\}$ otherwise. It seems that this advantage of many states is stronger than the disadvantage of less frequent backwards mutations to the root state which are lucky for ancestral reconstruction.

In summary, when $q \leq b$ the performance of the Fitch method is as poor as randomly picking a state. Only when
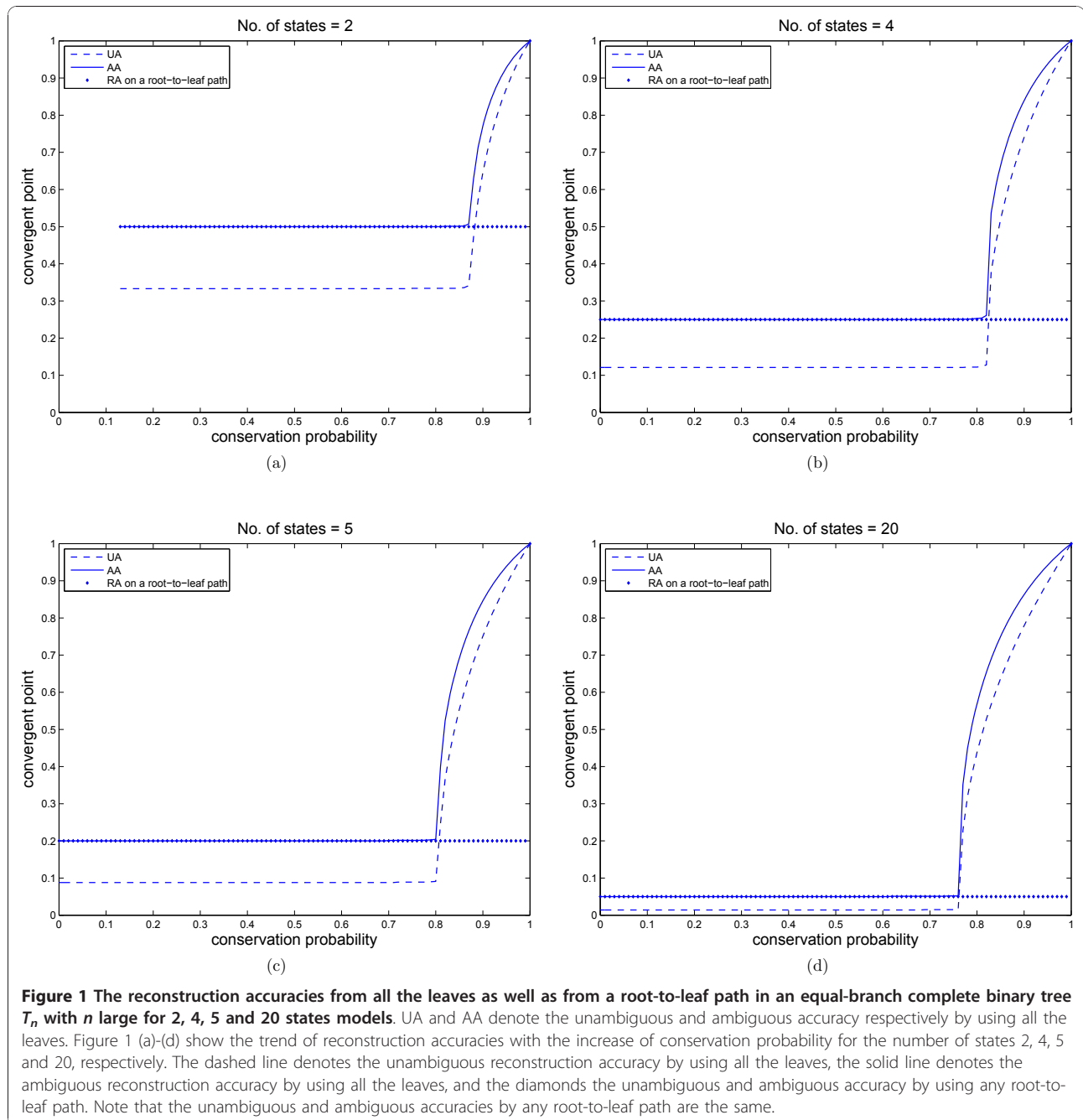
Figure 1 The reconstruction accuracies from all the leaves as well as from a root-to-leaf path in an equal-branch complete binary tree T_n **with** n **large for 2, 4, 5 and 20 states models**. UA and AA denote the unambiguous and ambiguous accuracy respectively by using all the leaves. Figure 1 (a)-(d) show the trend of reconstruction accuracies with the increase of conservation probability for the number of states 2, 4, 5 and 20, respectively. The dashed line denotes the unambiguous reconstruction accuracy by using all the leaves, the solid line denotes the ambiguous reconstruction accuracy by using all the leaves, and the diamonds the unambiguous and ambiguous accuracy by using any root-to-leaf path. Note that the unambiguous and ambiguous accuracies by any root-to-leaf path are the same.

Table 1 The estimated values of b **for the number of states** $N = 2, ..., 25$

			Estimated values of b					
N	2	3	4	5	10	15	20	25
b	0.875	0.839	0.821	0.809	0.784	0.774	0.768	0.763

Table 1 lists the estimated value of b by using Matlab. For convenience, we only show the values for $N = 2, 3, 4, 5, 10, 15, 20$ and 25.

$q > b$, the Fitch method could be used to reconstruct ancestral states and the performance improves quickly with the increase of q. As a conclusion, conservation probability is the most important factor to determine the performance of the Fitch method. The method is reliable only when q is large, which indicates that the taxa are highly similar. However, as we know, when taxa are not similar, no reconstruction method performs good, so more effort

should be made in developing a reliable method in this scenario. A suggestion for ancestral reconstruction is that, instead of treating all taxa as a whole, one first reconstructs subset of very similar taxa and make use of the reconstructed ancestral sequences to infer the ancestor of the whole taxa set.

Reconstruction Accuracies on Comb-shaped Trees

Due to their extreme unbalanceness, comb-shaped trees (or rooted caterpillars) are widely studied. A comb-shaped tree is a rooted binary tree where every internal node is adjacent to at least one leaf node. We studied two comb-shaped trees as shown in Figure 2. The tree in Figure 2 is called *equal-branch comb-shaped tree*, which assumes that substitution happens mostly at speciation events and thus each branch length can be considered as equal. The other tree is called *Hennigian comb-shaped tree*, which assumes that substitution occurs continuously during the course of evolution. More precisely, we assume that all interior edges have the same length and all leaves are equally far away from the root.

Equal-branch Comb-shaped Trees

In this tree, a descendant leaf X of the root is closer in evolutionary distance than other leaves. Let the substitution probability of any kind along any branch be p. Since the left subtree of the root is a branch and the right subtree is a repeat of the tree with one leaf less, the recurrence formula can be obtained by substituting $A_{i_X}^Z = A_i^n, A_i^Y = A_i^{n-1}$ for $i = 1, \cdots, 2N-1, A_1^X = 1$ and $A_j^X = 0$ for $j \neq 1$. We simulated the obtained system by Matlab and studied the asymptotic properties. The results for $N = 2, 4, 5$ and 20 are plotted in Figure 3.

An interesting observation is that, in contrast to complete binary trees, there is no interval of conservation probability such that AA converges to $\frac{1}{N}$ for any equal-branch comb-shaped tree. A possible reason is that the leaves that are close to the root dominate the reconstruction accuracy and their distances to the root do not increase with time. In addition, the limiting AA using the nearest root-to-leaf path is always greater than that using all taxa if conservation probability is larger than $\frac{1}{N}$.
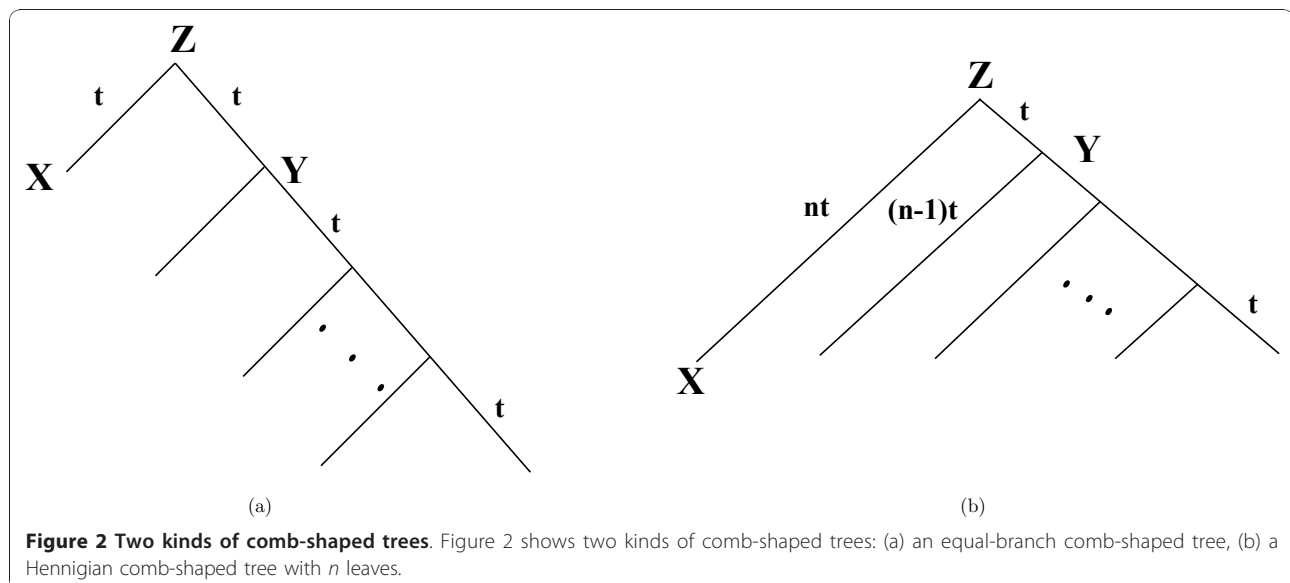
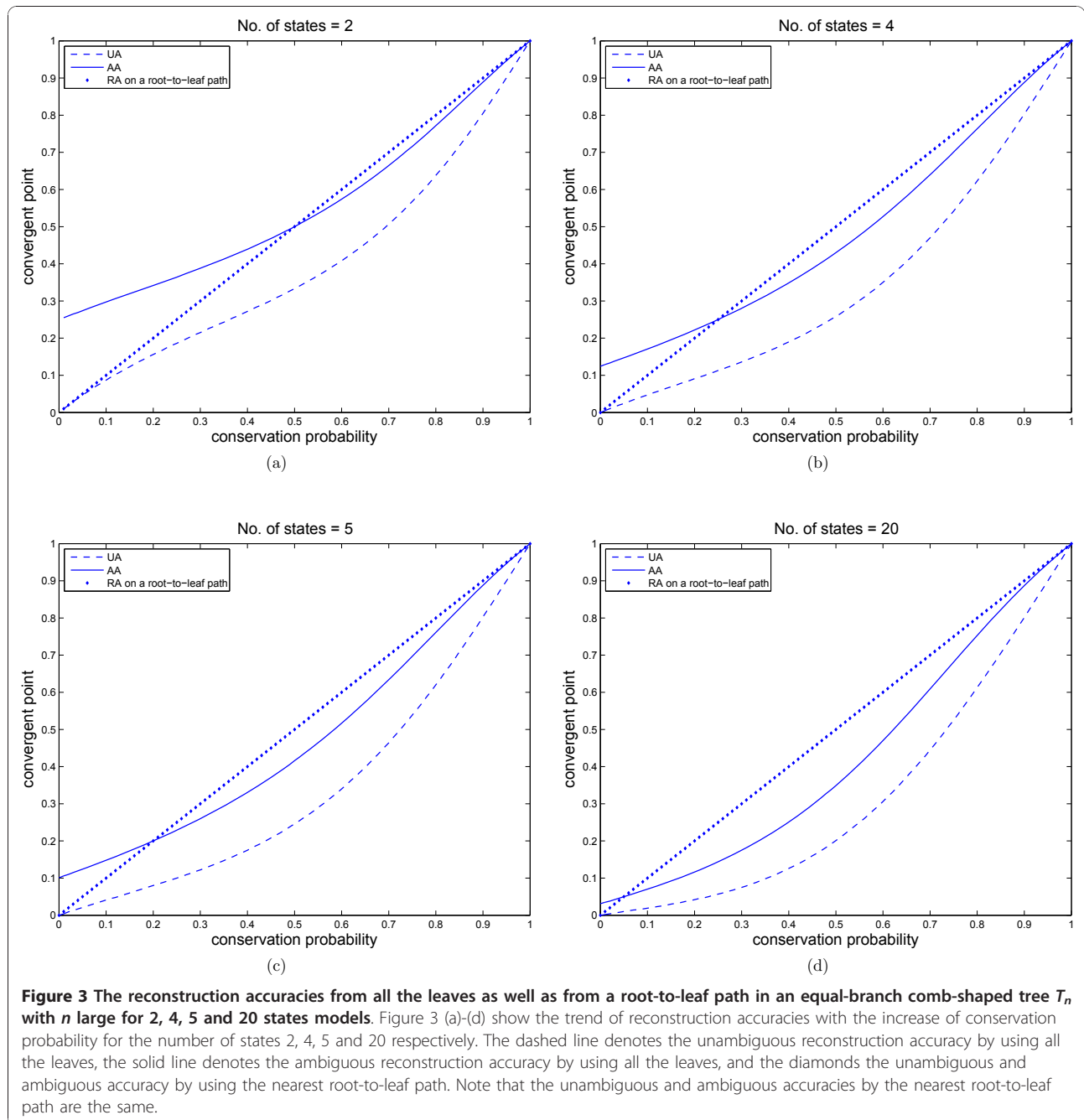#### *Hennigian Comb-shaped Trees*

Clearly, complete binary tree and Hennigian comb-shaped tree are the two kinds of extremal ultrametric trees. By comparing the reconstruction accuracies on both trees with those on real evolutionary trees, one can examine which extremal trees are more realistic. Under the Jukes-Cantor model, for a Hennigian comb tree in which each branch has its own length l, the substitution probability is $p = \frac{1}{N} - \frac{1}{N} e^{-N\lambda l}$, and the conservation probability is $q = 1 - (N-1)p = \frac{1}{N} + \frac{N-1}{N} e^{-N\lambda l}$, where λ is the substitution rate. Similarly, the recurrence system to calculate UA of the Fitch algorithm along the Hennigian tree can be derived from the general recurrence relations.

An observation from the recurrence system is that AA is always convergent to $\frac{1}{N}$. As an indication, the Fitch algorithm is not suitable for ancestral reconstruction on evolutionary trees with shape similar to Hennigian comb-shaped trees with many taxa. In addition, UA is

convergent to $\dfrac{N^{N-2}}{\sum_{i=l}^{N} N^{N-i} \dfrac{(N-1)!}{(N-x)!}}$ (See Additional File 3

for details). To illustrate the result, we also listed in Table 2

Figure 2 Two kinds of comb-shaped trees. Figure 2 shows two kinds of comb-shaped trees: (a) an equal-branch comb-shaped tree, (b) a Hennigian comb-shaped tree with *n* leaves.

Figure 3 The reconstruction accuracies from all the leaves as well as from a root-to-leaf path in an equal-branch comb-shaped tree T_n with n large for 2, 4, 5 and 20 states models. Figure 3 (a)-(d) show the trend of reconstruction accuracies with the increase of conservation probability for the number of states 2, 4, 5 and 20 respectively. The dashed line denotes the unambiguous reconstruction accuracy by using all the leaves, the solid line denotes the ambiguous reconstruction accuracy by using all the leaves, and the diamonds the unambiguous and ambiguous accuracy by using the nearest root-to-leaf path. Note that the unambiguous and ambiguous accuracies by the nearest root-to-leaf path are the same.

the simulated convergent values of UA from the recurrence system for some numbers N between 2 and 50.

Reconstruction Accuracies for Yule Trees with 1024 Leaves

As can be seen from Figure 4, the 1000 randomly generated Yule trees show the same qualitative behavior as the equal-branch complete binary tree with the same number of leaves. That is, there is an interval $[0, b]$ of

conservation probability q in which AA is almost equal to the probability of randomly picking a state, when $q > b$, the reconstruction accuracies are improved quickly with the increase of q. A possible explanation is that the number of taxa increases exponentially with the time, and correspondingly, very long edges are rare. Another interesting observation is that sometimes the reconstruction accuracies on the Yule trees are even higher than those of the complete binary tree. This contradicts

Table 2 The estimated values for UA of Hennigian comb-shaped trees when *n* is large for the number of states *N* = 2, ..., 50

	Estimated values of the limiting UA for Hennigian trees									
N	2	3	4	5	10	15	20	30	40	50
UA	0.3333	0.1765	0.1126	0.0797	0.0274	0.0145	0.0098	0.0053	0.0034	0.0022

Table 2 lists the estimated value of UA of Hennigian trees when *n* is large by using Matlab. For convenience, we only show the values for *N* = 2, 3, 4, 5, 10, 15, 20, 30, 40 and 50.

the intuition that equal-branch complete binary trees always maximize the reconstruction accuracies and leaves an open problem which trees do so.

Conclusions

In this paper, we study the unambiguous and ambiguous reconstruction accuracy of the Fitch method. We first present a general recurrence system as well as a program for calculating reconstruction accuracies on arbitrary trees. Based on the system and program, we analyze 3 special trees under the Jukes-Cantor evolutionary model, namely equal-branch complete binary trees, equal-branch comb-shaped trees, and Hennigian comb-shaped trees, as well as 1000 randomly generated Yule trees to simulate real evolutionary scenarios. From the analyses, we conclude that (1) for equal-branch complete binary trees, there always exists an interval [0, *b*] of conservation probability, in which the ambiguous reconstruction probability converges to $\frac{1}{N}$, the probability of randomly picking a state, when the

conservation probability is greater than *b*, both reconstruction accuracies increase rapidly, The randomly generated Yule trees also exhibit the same behavior, (2) For unbalanced trees like comb-shaped trees, the reconstruction accuracies using the nearest root-to-leaf path are always greater than or equal to those using all taxa. As a conclusion, more taxa are suggested for ancestral reconstruction when the tree topology is balanced and the sequences of taxa are highly similar, and a few taxa close to the root are recommended otherwise.

Availability and Requirements

The software as well as the source code in C_{++} to calculate the reconstruction accuracy of the Fitch method on any tree with arbitrary states under the one parameter Jukes-Cantor model can be found in Additional File 4. The reader is referred to the "install.txt" and "help.txt" file for the installation and usage of the program, or alternatively run the bash file "accuracy.out" in a Unix/
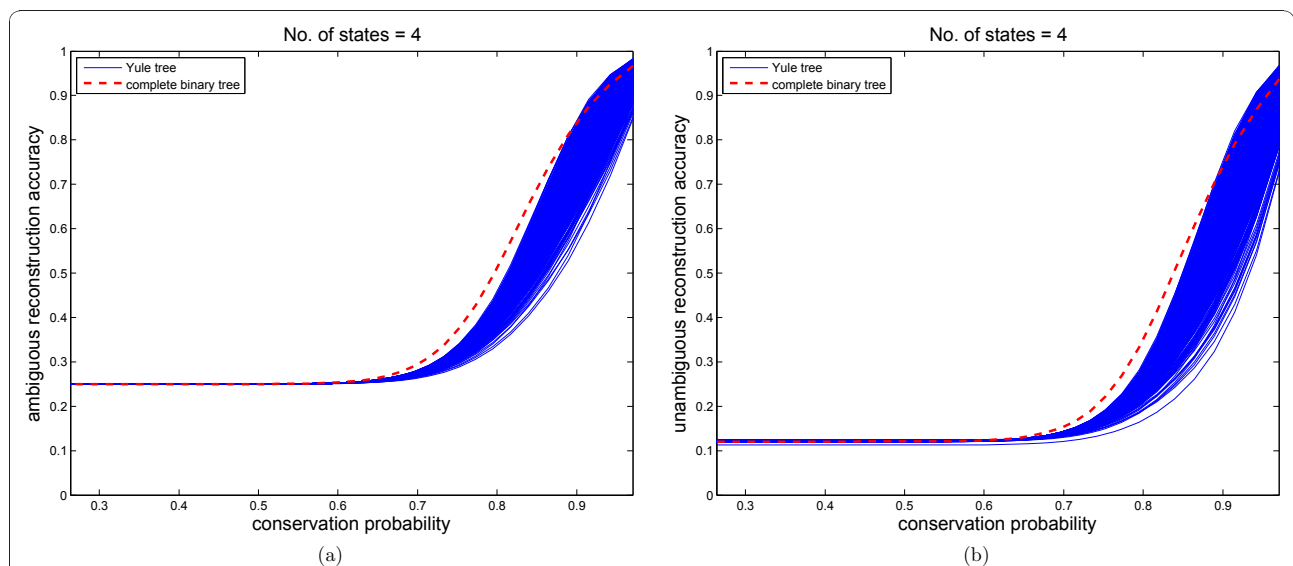

Figure 4 The comparison of unambiguous and ambiguous reconstruction accuracy between 1000 randomly generated Yule trees and the equal-branch complete binary tree with 1024 leaves. Figure 4 (a)-(b) show the trend of ambiguous and unambiguous reconstruction accuracies with the increase of conservation probability for 1000 randomly generated Yule trees and the equal-branch complete binary tree with 1024 leaves. The dashed line denotes the trend of reconstruction accuracy on the equal-branch complete binary tree, the solid lines denote the trend of reconstruction accuracy on the randomly generated Yule trees.

Linux system. The programs to draw the figures and tables are written in Matlab, which can also be found in Additional File 4. So Matlab should be installed to run these codes.

Additional material

Additional file 1: The recurrence system for the reconstruction accuracy of the Fitch method on *N*-state models. In this file, we provide the general recurrence system and initial conditions for calculating the reconstruction accuracy of the Fitch method on *N*-state models.

Additional file 2: The recurrence system of calculating reconstruction accuracies for the complete binary tree T_n. In this file, we provide the recurrence system and initial conditions for calculating the reconstruction accuracy of the Fitch method on the complete binary tree T_n with 2^n leaves.

Additional file 3: A sketch of the proof for the formula of limiting UA on Hennigian comb-shaped trees. In this file, we provide a sketch of the proof that the limiting UA on Hennigian comb-shaped trees with

N leaves is $\dfrac{N^N}{\sum_{i=l}^{N} N^{N-i} \dfrac{N!}{(N-x)!}}$.

Additional file 4: Programs and source codes to calculate the reconstruction accuracy and draw figures and tables. In this file, we present the software as well as its source code in C_{++} to calculate the reconstruction accuracy of the Fitch method on any tree with arbitrary states under the one parameter Jukes-Cantor model. The programs in Matlab to draw the figures and tables for extremal trees are also provided.

Authors' contributions

SG and JY conceived the methods. JL and JY performed the experiments. JY prepared the manuscript. All authors contributed to the discussion and have approved the final manuscript.

References

1. Pauling L, Zuckerkandl E: **Chemical paleogenetics: molecular restoration studies of extinct forms of lives.** *Acta Chem Scand* 1963, **17**:S9-S16.
2. Hills D, Huelsenbeck J, Cunningham C: **Application and accuracy of molecular phylogenies.** *Science* 1994, **264**:671-677.
3. Jermann T, Opitz J, Stackhouse S, Benner S: **Reconstructing the evolutionary history of the artiodactyl ribonuclease superfamily.** *Nature* 1995, **374**:57-59.
4. Zhang J, Rosenberg H: **Complementary advantageous substitutions in the evolution of an antiviral RNase of higher primates.** *Proc Natl Acad Sci* 2002, **99**:5486-5491.
5. Zhang C, Zhang M, Ju J, Neitfeldt J, Wise Jea: **Genome diversification in phylogenetic lineages I and II of Listeria monocytogenes: Identification of segments unique to lineage II populations.** *J Bacteriol* 2003, **185**:5573-5584.
6. Felsenstein J: *Inferring phylogenies* Sunderland, Massachusetts: Sinauer Associates; 2004.
7. Krishnan N, Seligmann H, Stewart C, De Koning A, Pollock D: **Ancestral sequence reconstruction in primate mitochondrial DNA: Compositional bias and effect on functional inference.** *Mol Biol Evol* 2004, **21**:1871-1883.
8. Tauberberger J, Reid A, Louren R, Wang R, Jin G: **Characterization of the 1981 influenza virus polymerase genes.** *Nature* 2005, **437**:889-893.
9. Clemente J, Ikeo K, Valiente G, Gojobori T: **Optimized ancestral state reconstruction using Sankoff parsimony.** *BMC Bioinformatics* 2009, **10**:51.
10. Crisp M, Cook L: **Do early branching lineages signify ancestral traits?** *Trends in Ecol Evol* 2005, **20**:122-128.
11. Fitch W: **Toward defining the course of evolution: minimum change for a specific tree topology.** *Syst Zool* 1971, **20**:406-416.
12. Sankoff D: **Minimal mutation trees of sequences.** *SIAM J Appl Math* 1975, **28**:35-42.
13. Sankoff D, Rousseau P: **Locating the vertices of a Steiner tree in an arbitrary metric space.** *Math Program* 1975, **9**:240-246.
14. Albert V: *Parsimony, Phylogeny, and Genomics* Natural History Museum, University of Oslo, Norway: Oxford Scholarship Online; 2007.
15. Yang Z, Kumar S, Nei M: **A new method of inference of ancestral nucleotide and amino acid sequences.** *Genetics* 1995, **141**:1641-1650.
16. Koshi J, Goldstein R: **Probabilistic reconstruction of ancestral protein sequences.** *J Mol Evol* 1996, **42**:313-320.
17. Huelsenbeck J, Bollback J: **Empirical and hierarchical Bayesian estimation of ancestral states.** *Syst Bio* 2001, **50**:351-366.
18. Huelsenbeck J, Ronquist F: **MRBAYES: Bayesian inference of phylogenetic trees.** *Bioinformatics* 2001, **17**:754-755.
19. Lio P, Goldman N: **Models of molecular evolution and phylogeny.** *Genome Res* 1998, **8**:1233-1244.
20. Zhang J, Nei M: **Accuracies of ancestral amino acid sequences inferred by parsimony, likelihood, and distance methods.** *J Mol Evol* 1997, **44**:139-146.
21. Salisbury B, Kim J: **Ancestral state estimation and taxon sampling density.** *Syst Biol* 2001, **50**:557-564.
22. Steel M: **Distributions on bicoloured evolutionary trees.** *PhD thesis* Massey University, New Zealand; 1989.
23. Maddison W: **Calculating the probability distributions of ancestral states reconstructed by parsimony on phylogenetic trees.** *Syst Biol* 1995, **44**:474-481.
24. Li G, Steel M, Zhang L: **More taxa are not necessarily better for the reconstruction of ancestral character states.** *Syst Biol* 2008, **57**:647-653.
25. Yang J: **Three mathematical issues in reconstructing ancestral genome.** *PhD thesis* National University of Singapore, Singapore; 2008.
26. Fischer M, Thatte B: **Maximum parsimony on subsets of taxa.** *J Theoret Bio* 2009, **260**:290-293.
27. Zhang L, Shen J, Yang J, Li G: **Analyzing the Fitch method for reconstructing ancestral states on ultrametric phylogenetic trees.** *Bull Math Bio* 2010, **72**:1760-1782.
28. Oakley T, Gu Z, Abouheif E, Patel N, Li W: **Comparative methods for the analysis of gene-expression evolution: An example using yeast functional genomic data.** *Mol Biol Evol* 2005, **22**:40-50.
29. Maddison W, Maddison D: **Mesquite: a modular system for evolutionary analysis.** 2010 [http://mesquiteproject.org].
30. Tuffley C, Steel M: **Links between maximum likelihood and maximum parsimony under a simple model of site substitution.** *Bull Math Bio* 1997, **59**:581-607.
31. Saitou N, Nei M: **The neighbor-joining method: a new method for reconstructing phylogenetic trees.** *Mol Biol Evol* 1987, **4**:406-425.
32. Lundstrom R, Tavare S, Ward R: **Estimating substitution rates from molecular data using the coalescent.** *Proc Natl Acad Sci* 1992, **89**:5961-5965.