HCPC: A New Parsimonious Clustering Method based on Hierarchical Characters for Morphological Phylogenetic Reconstruction

Hongwei Feng
Department of Information Science and Technology, Northwest University
Meng Liu
Department of Information Science and Technology, Northwest University
Bei Wang
Department of Information Science and Technology, Northwest University
Jun Feng (fengjun@nwu.edu.cn)
Department of Information Science and Technology, Northwest University
Jian Han
Early Life Institute, State Key Laboratory of Continental Dynamics, Department of Geology, Northwest University
Jianni Liu
Early Life Institute, State Key Laboratory of Continental Dynamics, Department of Geology, Northwest University

Research Article

Keywords: morphological data,phylogenetic reconstruction, hierarchical clustering, inapplicable data

Posted Date: January 6th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-138730/v1

HCPC: A New Parsimonious Clustering Method based on Hierarchical Characters for Morphological Phylogenetic Reconstruction

Hongwei Feng[1], Meng Liu[1], Bei Wang[1],Jun Feng[1,*], Jian Han[2], Jianni Liu[2]

[1] Department of Information Science and Technology, Northwest University, Xi'an 710127, China; hwfeng@nwu.edu.cn (H.F.); liumeng@stumail.nwu.edu.cn (M.L.);wangbei@stumail.nwu.edu.cn (B.W.);fengjun@nwu.edu.cn (J.F.)

[2] Early Life Institute, State Key Laboratory of Continental Dynamics, Department of Geology, Northwest University, Xi'an 710069, China;

elihanj@nwu.edu.cn (J.H.); eliljn@nwu.edu.cn (J.L.)

* Correspondence: Jun Feng[1]

Email address: fengjun@nwu.edu.cn

Abstract

Background: Phylogenetic trees are reconstructed frequently to provide a better interpretation of the evolutionary history of species. However, most traditional methods ignore the hierarchical relationships among characters and neglect the inapplicable state that frequently exists in the morphological data, resulting in poor performance of the phylogenetic analysis.

Results: In this study, we propose a phylogenetic clustering method based on hierarchical characters. Accordingly, we call our method Hierarchical Characters Parsimonious Clustering(HCPC). To combine prior phylogenetic knowledge and treat the inapplicable state more reasonably, two stages are proposed, i.e., Phylogenetic reconstruction and parsimonious tree search. During phylogenetic reconstruction, HCPC is able to infer the shared ancestral relationships among species. For the search of the parsimonious tree, we use a simulated annealing algorithm to heuristically search the phylogenetic tree based on the parsimony criterion. In addition, HCPC

combines asymmetric binary relationships and character hierarchies to solve the problem of the ambiguity of the inapplicable state.

Conclusion: The experimental results show that the proposed method provides better performance of phylogenetic analysis than existing methods and a scientific and quantitative basis for biologists to study species evolution.

Keywords: morphological data, phylogenetic reconstruction, hierarchical clustering, inapplicable data

Introduction

Phylogenetic reconstruction is an interesting topic in biology and an important analysis method to investigate the evolutionary process and to develop an understanding of when and where speciation events may have occurred (Yang et al., 2012). A phylogenetic tree consists of edges, internal nodes, and leaves. The leaves represent the Operational Taxonomic Units (OTUs), which are the actual species used to reconstruct a tree. The internal nodes are the Hypothetical Taxonomic Units (HTUs), which represent the hypothetical shared ancestors to all other species arising from them. The species arising from an internal node are the direct descendants of this internal node. The leaves connected to the same internal node are called a sister group and have the closest shared ancestral relationship (Scott et al., 2012). The edges represent the relationship between the two nodes.

Traditionally, there are two types of methods for phylogenetic reconstruction: distance-based methods and optimization-based methods (Goloboff et al., 2005; Yang et al., 2012). In a general way, distance-based methods depend on various types of data to reconstruct a tree, i.e., the genetic distance of the sequences and the Euclidean distance (Felsenstein, 1988; Scott et al., 2012). In terms of distance-based methods, there are commonly used algorithms, i.e., the Unweighted Pair Group Method with Arithmetic Mean (UPGMA) (Sneath et al., 1973), Neighbor-Joining (NJ)

(Saitou et al., 1987; Tamura et al., 2004), and the Fitch-Margoliash method (Farabee, 2007). Each algorithm possesses known properties so that the reconstructed tree is very similar to the real tree (Scott et al., 2012). However, optimization-based methods rely on a variety of phylogenetic characters including genetic, morphological, behavioral, and molecular to perform the analysis (Goloboff et al., 2005). Generally, these algorithms such as Maximum Parsimony (MP), Maximum Likelihood (ML), and Bayesian Inference (BI) are based on an optimization criterion (Lewis, 2001; Goloboff et al., 2005; O'Reilly et al., 2016; Puttick et al., 2017).

However, existing methods consider characters to be logically independent and ignore the hierarchical relationships among them (Zaragüetabagils et al., 2007; Nelson et al., 2010). If one character state depends on another character state in a species, it means that there is a hierarchical relationship between the characters (Lee et al., 1999; Seitz et al., 2000). The hierarchical relationship indicates that characters in phylogenetic analysis have some logical dependencies because complex biomorphologic characters can be broken down into secondary characters with logical relationships. These decomposable complex characters are called upper characters, and the secondary characters decomposed by complex characters are called lower characters. The hierarchical relationship among characters implies that the upper character and its lower character are logically dependent. In other words, the lower character is only applicable to a species that has its upper character (Seitz et al., 2000). For example, suppose some species have no tails, some have blue tails, and others have red tails. The tail color character and tail length character depend on the presence of tail character. The hierarchical relationship between the characters in this example is shown in Fig. 1. The lower characters, such as tail color and tail length are applicable only for species that have a tail. The species with no tails are recorded as inapplicable state in tail color character (Strong et al., 2010; Wilkinson, 2010).

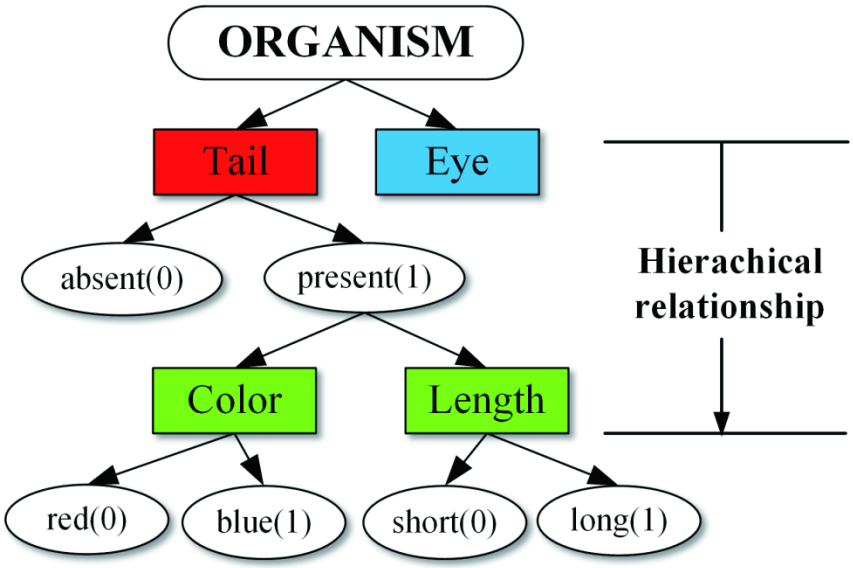


Fig. 1. Example of the hierarchical relationship among characters. The rectangles represent the characters and the ellipses represent the states of the character. The red, green and blue rectangles represent the upper character, lower character, and solitary character, respectively.

As is well known, the treatment of inapplicable state is a concern in phylogenetic reconstruction, especially for paleontological research, in which morphological data is the only material for phylogenetic reconstruction (Zaragüetabagils et al., 2007). Unfortunately, both distance-based methods and optimization-based methods are problematic when some inapplicable states exist in the morphological data (Zaragüetabagils et al., 2007; Goloboff et al., 2017; O'Reilly et al., 2018). Meanwhile, there are two main approaches to treat inapplicable state. One frequently used method is Missing Data Replacement (MDR), which treats inapplicable state as missing data. However, this result in an assumption of homology where no homology exists (Maddison, 1993; Seitz et al., 2000; Platnick et al., 2010). The second is Separate Value Replacement (SVR), which treats the inapplicable states as separate character states; this approach is often used in the MP method. However, the inapplicable state is not comparable to other states defined by the characters and this method implicitly weights the upper characters of the species (Lee et al., 1999; Seitz et al., 2000). In other words, all the aforementioned methods often result in poor performance of phylogenetic

analysis (Congreve et al., 2016). Different from other character states, the inapplicable state does not imply that a certain homology hypothesis can be used to reconstruct phylogenetic trees (Lee et al., 1999), whereas the above-mentioned method assumes that the inapplicable states implies an assumption of homology.

In the morphological data, the hierarchical relationship leads to inapplicable state. Furthermore, when manually reconstructing phylogenetic trees, the species with the shared derived character states can be distinguished from other species based on the polarity of the characters (Huang, 1996). Thus, it is crucial to explore the hierarchical relationships and the polarity of the characters when reconstructing phylogenetic trees. Therefore, we divide the framework of HCPC into two steps, namely, phylogenetic reconstruction based on the hierarchical relationship and a parsimonious tree search. HCPC reconstructs phylogenetic trees based on the asymmetric binary relationships and hierarchical architecture of characters. In particular, the polarity of characters is quantified into a distance calculation to measure the shared ancestral relationship among species, and the character vectors of the internal nodes are inferred from the hierarchical relationships. Therefore, when reconstructing a phylogenetic tree, no homology hypothesis regarding inapplicable states is required. Our experimental results show that in most cases, HCPC achieves better performance than traditional approaches such as MDR and SVR in terms of the Robinson-Foulds (RF) distance.

The paper is organized as follows: Section 2 introduces the framework of HCPC. We present the experimental results and the analysis in Section 3. Finally, Section 4 describes the conclusions drawn from the work.

Method

It is well known that if species share more derived character states, they may have closer shared ancestral relationships (Huang, 1996). Moreover, in phylogenetic analyses, inapplicable states are attributed to the hierarchical relationships among characters (Lee et al., 1999). Therefore, when

inapplicable states exist in the morphological data, HCPC combines the polarity of the characters with the hierarchical relationships among the characters to reconstruct a phylogenetic tree. Specifically, HCPC reconstructs the phylogenetic trees using phylogenetic clustering and searches for the parsimonious trees using a simulated annealing algorithm. Fig. 2 shows the framework of the proposed HCPC method.

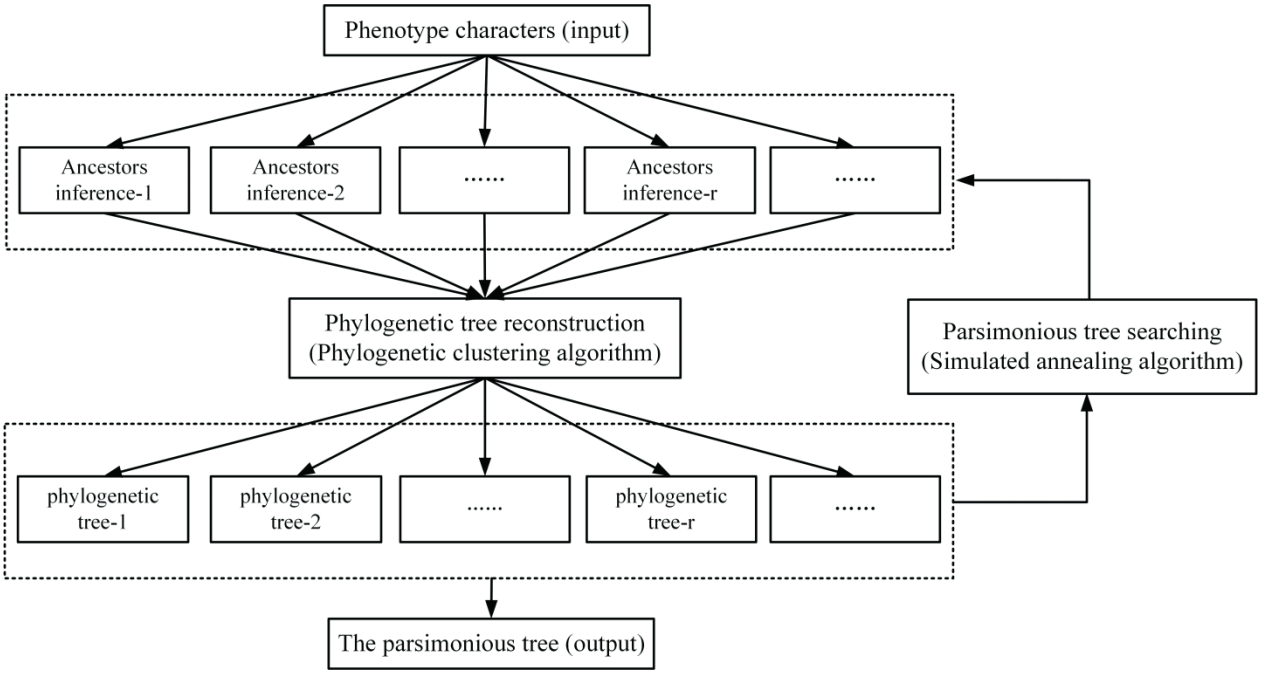


Fig. 2. The framework of phylogenetic analysis method based on HCPC.

2.1 Phylogenetic clustering for phylogenetic reconstruction

When reconstructing phylogenetic trees, the shared ancestral relationships among the species are inferred based on the shared derived index. The species with the maximum value of the shared derived index are selected to generate an internal node. Then the character vector of this internal node is inferred and added to the data matrix D as a "new" species instead of its direct descendants. The shared ancestral relationships among species at different levels are established after repeating this process. The flowchart of the phylogenetic clustering based on the character hierarchy for phylogenetic reconstruction is shown in Fig. 3.

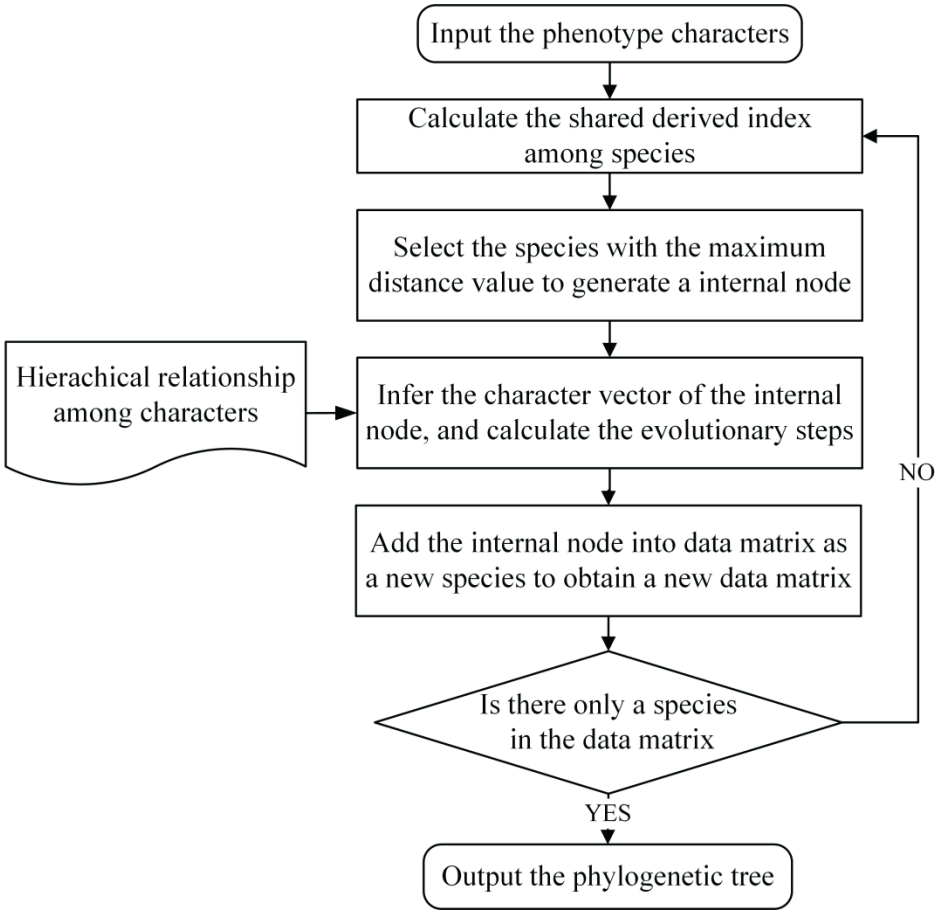


Fig. 3. Flowchart of the phylogenetic clustering based on the character hierarchy for the phylogenetic reconstruction.

2.1.1 Measuring the shared ancestral relationship using the shared derived index

Let $D\{X_1, \dots, X_n\}$ be a finite character matrix of n species where X_i represents the i-th species' character states. The state of the p-th character for X_i is denoted as x_{ip}. The number of characters used to reconstruct phylogenetic trees is denoted as m; therefore, the character vector of species X_i is $\{x_{i1}, \dots, x_{im}\}$. Suppose that the upper character of the p-th character is the q-th character. That is, when some species lack the q-th character, the p-th character is inapplicable.

According to the polarity of characters, the character states can be divided into two categories, i.e., derived states and ancestral states. The polarity of characters is a dialectical relationship between the derived states and the ancestral states. Let s_p be the ancestral state of the p-th

character. Thus, $S\{s_1, \ldots, s_m\}$ marks the ancestral state of each character, which is an ancestor's inference. Thus, the character states in the character matrix that are same as the ancestor's inference are the ancestral states. Otherwise, the remaining states are derived states. Specifically, if $x_{ip} = s_p$, x_{ip} is an ancestral state for species X_i; otherwise, x_{ip} is a derived state for species X_i.

The derived state and the ancestral state are an asymmetric binary relationship for characters according to the principles of phylogenetic analysis. The number of identical derived character states between X_i and X_j is defined as a shared derived index $d(X_i, X_j)$, which measures the shared ancestral relationship between X_i and X_j. The shared derived index $d(X_i, X_j)$ is obtained as follows:

$$d(X_i, X_j) = \sum_{p=1}^{m} \left[(x_{ip} = x_{jp}) \cap \left((x_{ip} \neq s_p) \cup (x_{jp} \neq s_p) \right) \right], where\ x_{ip} \neq \text{"-"} \text{ and } x_{jp} \neq \text{"-"}. \quad (1)$$

For each species, the shared derived index is calculated in pairs. From this, we can obtain the metric matrix M_d defined in Eq. (2).

$$M_d = \begin{bmatrix} d(X_1, X_1) & \cdots & d(X_n, X_1) \\ \vdots & \ddots & \vdots \\ d(X_1, X_n) & \cdots & d(X_n, X_n) \end{bmatrix} \quad (2)$$

2.1.2 Inferring the character vectors using the hierarchical relationships

Suppose that $d(X_i, X_j)$ is the maximum value in the metric matrix M_d; therefore, X_i and X_j are selected to generate an internal node, a hypothetical species H is the hypothetical ancestor of X_i and X_j. Suppose the q-th character is the upper character of the p-th character. Then the character vector of H is inferred on the basis of the character vectors of X_i and X_j. The details of the calculation for $H\{x_1, \ldots, x_m\}$ are defined as follows:

$$H_p = \begin{cases} s_p, & \text{if } x_{ip} = s_p \text{ or } x_{jp} = s_p. \\ x_{ip}, \text{if } \begin{cases} x_{ip} = x_{jp}, x_{ip} \neq s_p, x_{ip} \neq " - ", x_{jp} \neq s_p, \text{and } x_{jp} \neq " - ". \\ x_{ip} = \text{"-"} \text{ or } x_{jp} = \text{"-"}, \text{and } H_q = x_{iq}. \end{cases} \\ x_{jp}, & \text{if } x_{ip} = \text{"-"} \text{ or } x_{jp} = \text{"-"}, \text{and } H_q = x_{jq}. \end{cases} \quad (3)$$

The character vector H_p of species H is inferred according to the ancestral state of the p-th character s_p. In particular, when the p-th character is an inapplicable state, the state of H_p is inferred by the state of the q-th character H_q. Therefore, the character vector of node H can be determined and the internal node H is added to the matrix D as a "new" species instead of X_i, X_j.

After generating a new hypothetical ancestor H, the difference between any two nodes X_i, X_j (including the HTUs) can be calculated by:

$$\mathrm{L}(X_i, X_j) = \sum_{k=1}^{m} |X_{ik} - X_{jk}| \quad (4)$$

Let T_s represents a phylogenetic tree. The tree length of T_s is denoted as $L(T_s)$. $L(T_s)$ is the sum of the character changes between all adjacent nodes.

2.1.3 An example to illustrate phylogenetic clustering

In this section, an example is given to illustrate the specific implementation process of phylogenetic clustering. As shown in Fig. 4, there are five characters in the data matrix. Character 1 and character 3 are solitary characters. Character 2 is an upper character so that character 4 and character 5 are applicable only when character 2 is "1". There are four species including $X_1\{000\text{--}\}, X_2\{101\text{--}\}, X_3\{01101\}$, and $X_4\{11111\}$(Fig. 4).

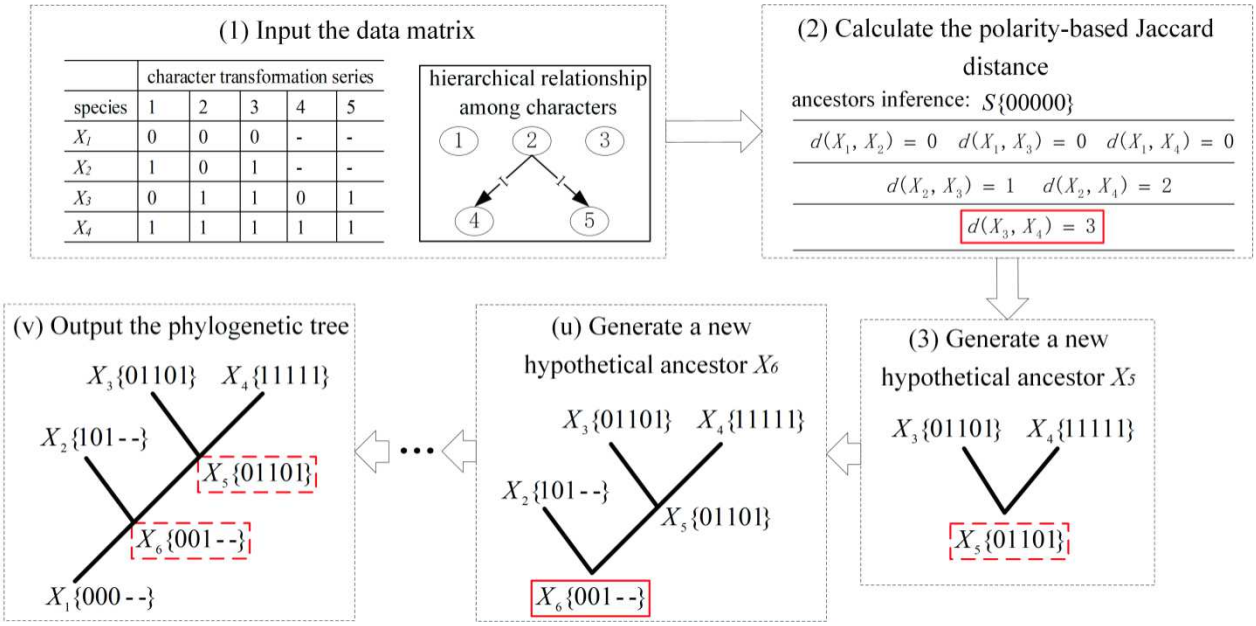


Fig. 4. Diagram of an example of Phylogenetic Reconstruction. The red solid line box is the maximum shared derived index and the red dotted boxes are the hypothetical species.

Suppose that the ancestor's inference is S{00000}, then the shared derived index is calculated based on the ancestor's inference S{00000}. X_3 and X_4 have the maximum value of shared derived index so that X_3 and X_4 are selected to generate a new hypothetical ancestor X_5. The character vectors of X_5 are inferred according to X_3 and X_4. Subsequently, X_5 is added to the data matrix as a "new" species instead of X_3 and X_4. Repeat the above steps, then X_2 and X_5 are selected to generate a new hypothetical ancestor X_6. There are some inapplicable states in the character vectors of X_2 so that the states of character 1, 2, and 3 for X_6 are inferred according to the polarity of character 1, 2, and 3, whereas the states of character 4 and character 5 for X_6 are inferred based on the state of character 2. The state of character 2 is "0"; therefore, the states of character 4 and character 5 are "-" for X_6. After inferring the character vectors, X_6 is added to the data matrix as a "new" species instead of X_2 and X_5. We repeat this process until the parsimonious tree has been reconstructed.

2.2 Simulated annealing algorithm for the parsimonious tree search

There are several methods for establishing the polarity of the characters, including outgroup comparison, stratigraphic positions, and developmental biology (Baum et al., 1996; Foote et al., 2007; Tian et al., 2017). However, none of these is an absolutely convincing method (Tian et al., 2017). Therefore, an exhaustive search is conducted on the ancestor's inference and various possible phylogenetic trees are calculated. However, it is an NP-hard problem to search for the parsimonious tree in the tree space. What should be of concern is that the tree with the minimum evolutionary steps to explain the change in the characters is the most parsimonious tree in the tree space based on parsimony principle (Desper et al., 2002; Farris, 2010). In other words, the tree with the shortest tree length is the most parsimonious tree and represents the object of interest here. Therefore, a simulated annealing algorithm (SA) is applied to search for the parsimonious tree heuristically according to the principle of parsimony.

The SA is a stochastic optimization algorithm based on the Monte-Carlo iterative strategy. The SA with asymptotic convergence is a global optimization algorithm and has resulted in a probability of l in the search for the optimal solution (Mahapatra et al., 2018). Furthermore, the parsimonious tree obtained by the SA is independent of the initial solution.

In this study, the SA is decomposed into three parts: the ancestor's inference space, objective function, and initial ancestor's states. The ancestor's inference space consists of all possible states for each character. According to the parsimony principle, the parsimony score $\mathrm{L}(T_S)$ is the objective function. $\mathrm{S}\{00 \ldots 00\}$ is the initial ancestor's states, which consists of state "0" of each character. For each character, a random state ("0" or "1") is selected to obtain a new ancestor's inference S_r. The difference between $\mathrm{L}(T_S)$ and $\mathrm{L}(T_{S_r})$ is an attenuation factor. In the SA, the

inner loop termination criterion consists of a small change in the attenuation factor in the successive loops, whereas the outer loop termination criterion is to reach a certain number of iterations.

Experimental Results and Analysis

To evaluate the performance of HCPC, we conducted experiments on seven morphological datasets, including three datasets of living species and four paleontological datasets. In the studies by Bouamer et al. (2003), Tang et al. (2014), and Chen et al. (2015), datasets of living species were used and the model trees were obtained using state-of-art methods such as MP and BI. Yang et al. (2015), Liu et al. (2011), Han et al. (2017), and Liu et al. (2010) used paleontological morphological datasets. The model trees for the paleontological species are generated based on the opinions of paleontologists (Aberer et al., 2013). Paleontological datasets contain a large amount of inapplicable states, whereas datasets of living species contain few. Currently, HCPC is only appropriate for binary characters. Therefore, multistate characters have to be transformed into binary characters using binary coding.

We use the RF distance to assess the difference between the reconstructed phylogenetic trees; it determines the number of unshared portions between the inferred tree and the model tree (Goloboff et al., 2010). The RF distance calculation can be expressed by Eq. (8):

$$\mathrm{RF}(T_1, T_2) = \frac{|split(T_1)|+|split(T_2)|-2|split(T_1)|\cap|split(T_2)|}{2(n-3)} \tag{8}$$

where n is the number of species. $|split(\mathrm{tree})|$ is the number of split tree sets of the tree. T_1 represents the inferred tree, T_2 represents the model tree. The inferred tree refers to the evolutionary tree constructed with the proposed method and the model tree refers to the phylogenetic tree accepted by biologists. The range of the RF distance is $0 \leq \mathrm{RF} \leq 1$. Generally speaking, the shorter the RF distance between the inferred tree and the model tree, the more similar the trees are. (Robinson et al., 1981; Wilkinson, 2010).

3.1 Experimental results

Based on the characterization of the morphological datasets and the prior knowledge of biologists, a character hierarchy model is constructed. We use the paleozoic lobopodians of Yang et al. as an example (Fig. 5).

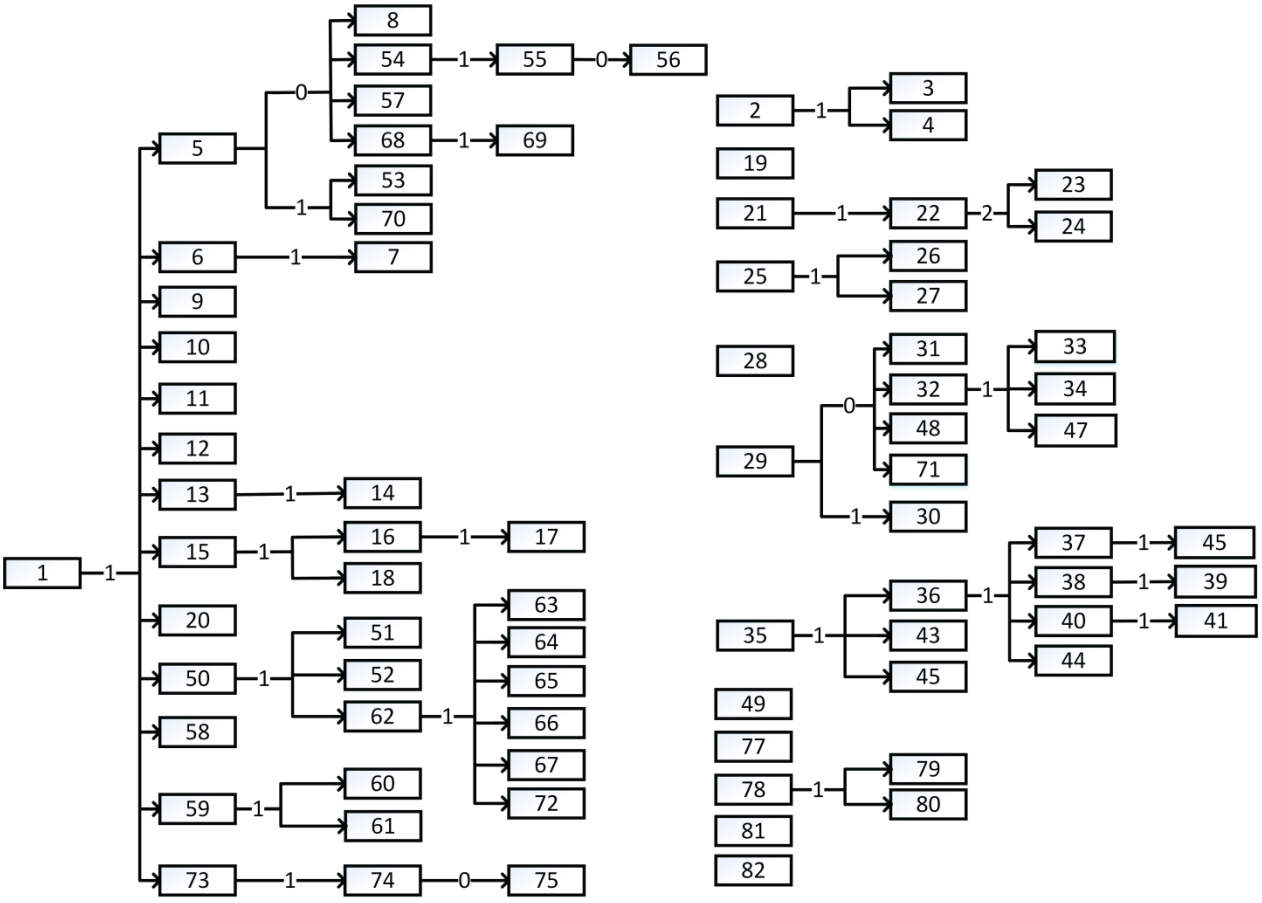
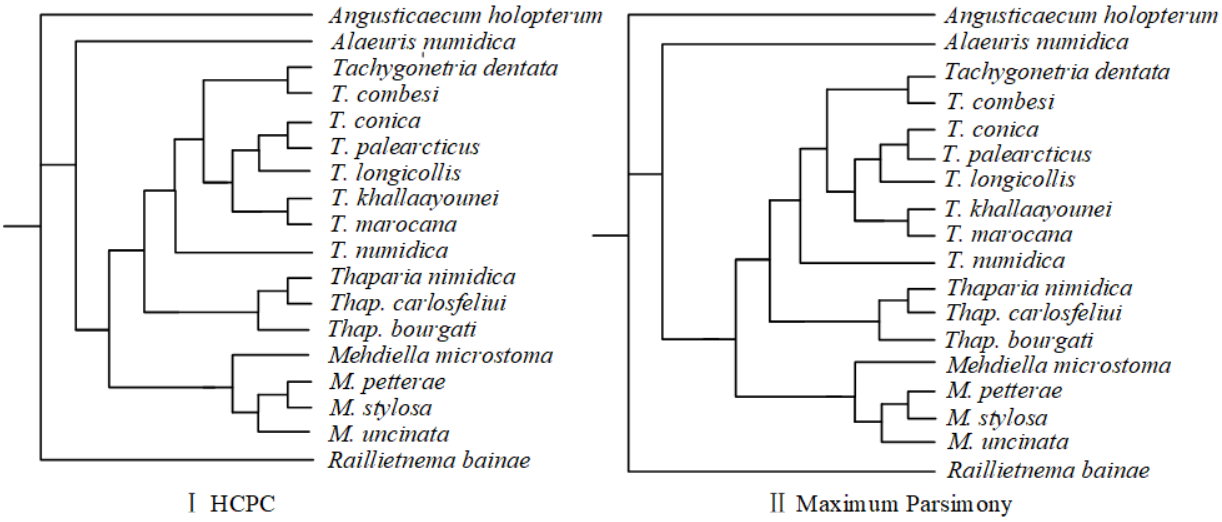


Fig. 5. A diagram of the hierarchical structure of characters for paleozoic lobopodians (part). The number in the box represents the No. character and the number on the arrow represents the character state.

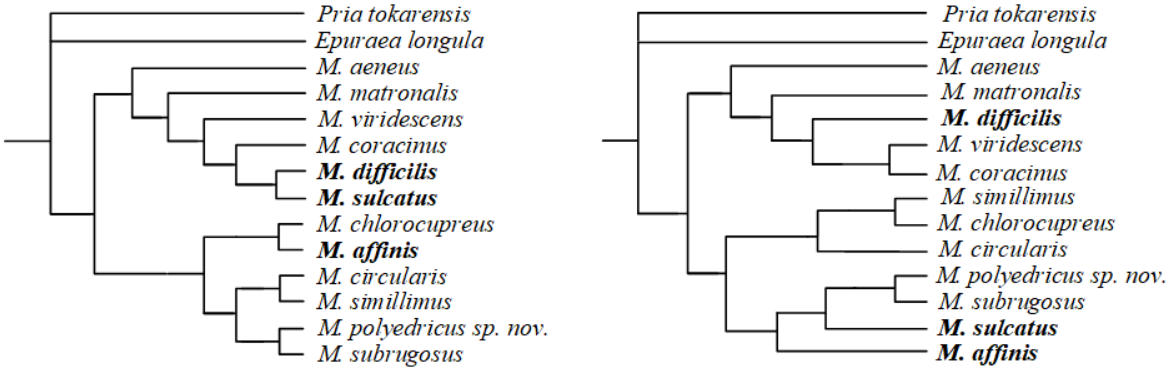
In Fig. 5, the character hierarchy model is a multi-level model. That is, multiple sub-characters may be nested under the feature. For example, when the state of character 21 (radially symmetrical circumoral structures) is "1", character 22 (nature of the circumoral structures) is a sub-character of character 21.

In order to verify the performance of the proposed method for evolution analysis, the proposed method is compared with the maximum parsimony method using the living species datasets (Bouamer et al., 2003; Tang et al., 2014; Lin, 2015) .



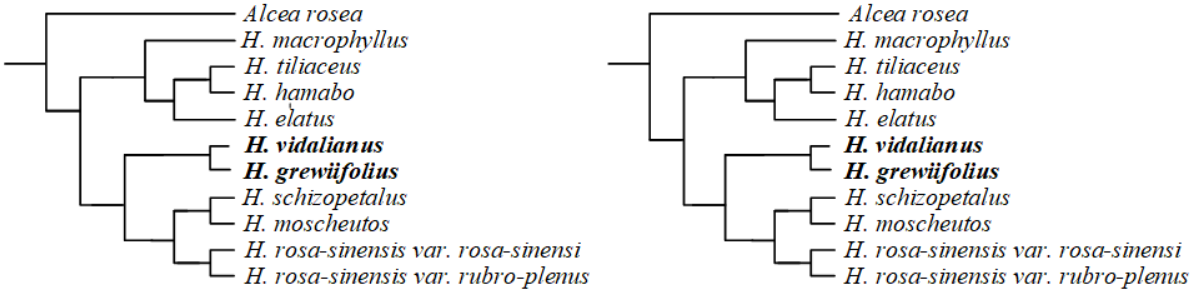
Ⅰ HCPC　　Ⅱ Maximum Parsimony

(a) Pharyngodonidae



Ⅰ HCPC　　Ⅱ Maximum Parsimony

(b) *Meligethes*



Ⅰ HCPC　　Ⅱ Maximum Parsimony

(c) *Hibiscus*

Fig. 6. A comparison of the phylogenetic trees reconstructed by HCPC and MP for living species. (A), (C), and (E) are the phylogenetic trees constructed using the HCPC method for the Bouamer & Morand dataset (Pharyngodonidae), the Chen et al. dataset (Meligethes), and the Tang et al. dataset (Hibiscus), respectively. (B), (D), and (F) are the phylogenetic trees constructed using the MP method for the Bouamer & Morand dataset (Pharyngodonidae), the Chen et al. dataset (Meligethes), and the Tang et al. dataset (Hibiscus), respectively. Species with different positions on the inferred trees and the model trees are marked in bold.

Fig. 6 shows that the inferred trees and the model trees are basically the same for the living species. In this study, we compare the NJ method with HCPC. Fig. 7 shows the performance of the HCPC and NJ methods for the three living species datasets. We observe that the inferred trees reconstructed by HCPC are more similar to the model trees than those based on the NJ method (Fig. 7).

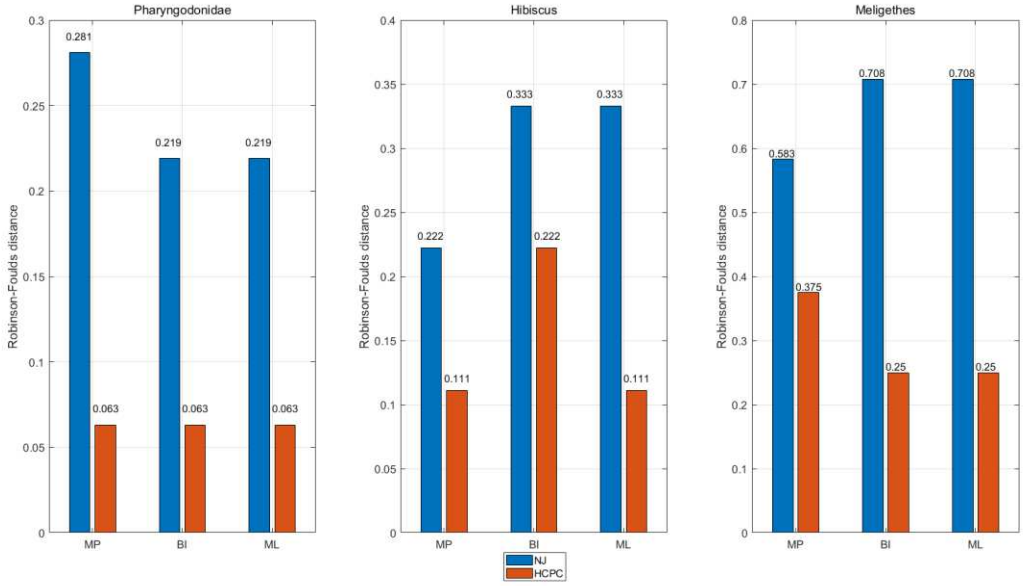


Fig. 7. RF distance for the NJ method and HCPC for living species.

We compare the HCPC method with MP-MDR, MP-SVR, ML, and BI. Fig. 8 shows the RF distance between the inferred tree and the model tree obtained by different methods for the paleontological datasets. Compared with the other methods, HCPC has a smaller RF distance for

the datasets of Yang et al. (2015), Liu et al. (2011), and Han et al. (2017). The HCPC method results in a decrease in the RF distance by 45.3% compared to the ML method for the dataset of Liu et al. (2011) and a decrease by 50.1% compared to the BI D for the dataset of Yang et al. (2015). The performance of the HCPC method is worse than that of the ML method for the dataset of Liu et al. (2010) (Fig. 8).

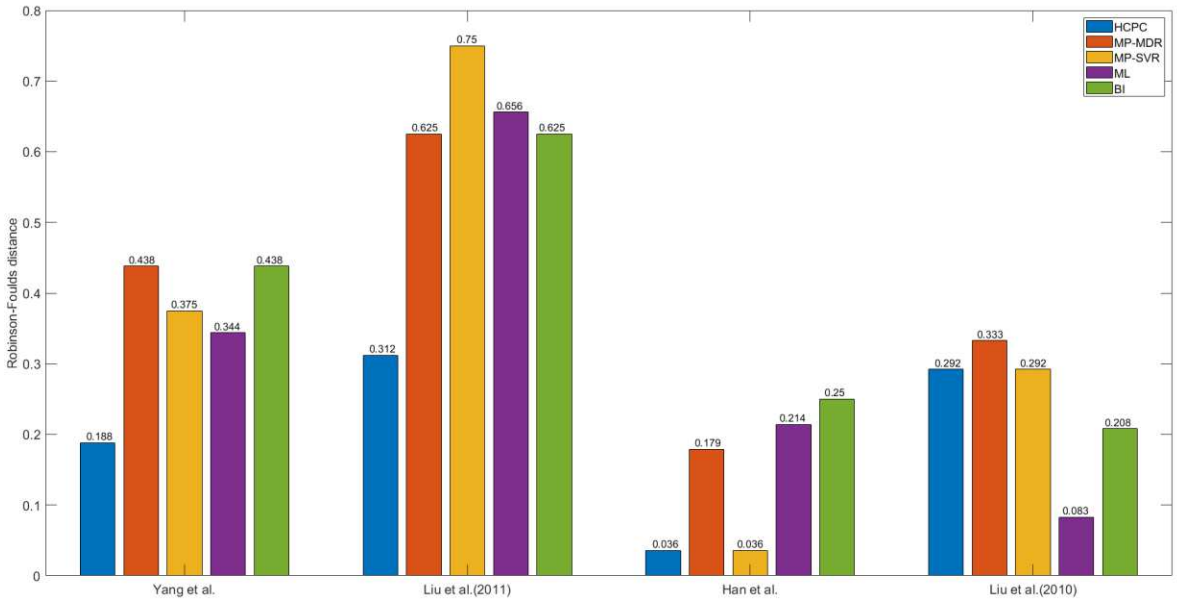


Fig. 8. RF distance for the HCPC and other methods for paleontological species. The MP-MDR, ML, and BI methods use the MDR to treat inapplicable states; the MP-SVR uses SVR to treat inapplicable states.

3.2 Discussion

In terms of the RF distance, HCPC outperforms the NJ method for morphology datasets of living species, indicating that HCPC results in a tree that is more similar to the actual phylogenetic tree. In phylogenetic analysis, the shared derived character states among species are the only evidence for shared ancestral relationships (Foote et al., 2007). Thus, the proposed phylogenetic clustering method measures the shared ancestral relationship among the species based on the shared derived index. Furthermore, under the parsimony criterion, no hypothesis is required regarding parallel evolution and reverse evolution when inferring the character vectors. Therefore, HCPC is a greedy

algorithm based on the principle of parsimony, so that HCPC always ensures that the current internal node has evolved its descendants with the least number of evolutionary steps when reconstructing a phylogenetic tree. That is, HCPC prefers that the similarity based on the parsimony assumption is interpreted as parallel evolution or evolutionary reversals. HCPC is effective when the evolution rate is low (the probability of parallel evolution and reverse evolution is also low) or the characters have evolved independently. To some extent, HCPC is also effective when the evolution rates of all species are equal.

Since HCPC is based on hierarchical characters for phylogenetic tree construction, HCPC outperforms the other methods when morphological data are used, especially paleontological morphology data. Thus, HCPC avoids explaining inapplicable states through the assumption of homology. Moreover, the character vectors of the internal nodes are inferred based on the polarity of the characters and the hierarchical relationships among the characters. However, MDR often results in an increase in missing data (Zaragüetabagils et al., 2007). For missing data, every state is possible, whereas, for inapplicable state is impossible (Platnick et al., 2010). Large amounts of missing data reduce the accuracy of the phylogenetic tree for traditional methods (Maddison, 1993). SVR results in extra emphasis on the absence of a single structure and acts as a form of weighting in MP; this approach often results in an inappropriate phylogenetic tree (Lee et al., 1999).

There are two reasons for the poor performance of HCPC for the Yang et al. (2015) dataset. Firstly, the problem of inapplicable states only occurs if two or more regions with applicable states in the branches are separated by regions with species with inapplicable character states (Maddison, 1993). Secondly, some distortions inevitably occur when determining the shared ancestral relationships among species; therefore, HCPC may fall into local optimum. However, in most cases, HCPC provides inferred trees that are very similar to the model trees and the main branches exhibit good consistency between the inferred trees and the model trees. The failure to determine a

sufficient number of characters for the Tang et al. (2014) dataset is an important reason for the low confidence in that particular phylogenetic tree. Furthermore, some individual diversity is lost when multivariable data are reduced to a single dimension to measure the shared ancestral relationship among species.

Conclusion

To address inapplicable states more elegantly, the proposed HCPC method incorporates additional prior phylogenetic knowledge including the polarity of the characters and the hierarchical relationships among characters to reconstruct the phylogenetic trees. In this way, the processing of inapplicable states does not involve an assumption of homology. The experimental results show that HCPC results in better performance for phylogenetic tree construction from morphological data than other common methods, especially for early paleontology. In future research, we will focus on using HCPC to obtain global optimal solutions at lower computational costs. In addition, we will also improve HCPC for multivariate character conversion.

Acknowledgments

The authors are particularly grateful to Mr. Qigao Jiangzuo for some useful comments on the manuscript, and the generous guidance of biology theory.

Funding

This work was sponsored by Northwest University Paleontological Bioinformatics Innovation Team(2019TD-012) and General program of Natural Science Foundation of Shaanxi Province (2019JM-494).

Abbreviations

HCPC:Hierarchical Characters Parsimonious Clustering; OTUs:Operational Taxonomic Units; HTUs:Hypothetical Taxonomic Units; UPGMA:the Unweighted Pair Group Method with

Arithmetic Mean; NJ:Neighbor-Joining; MP:Maximum Parsimony; ML:Maximum Likelihood; BI:Bayesian Inference; MDR:Missing Data Replacement; SVR:Separate Value Replacement; RF distance:Robinson-Foulds distance; SA:simulated annealing algorithm

Availability of data and materials

All morphological datasets analyzed in this paper come from published papers, which are availab le at https://github.com/utopfish/morphological-data

Ethics approval and consent to participate

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Consent for publication

Not applicable.

Authors' contributions

Conceptualization, H.F.; methodology, H.F. and M.L.; formal Analysis, M.L.; writing—original draft, H.F. and M.L.; writing—review and editing, M.L., B.W. and J.F.; funding acquisition, J.L. and J.H

Authors' information

Hongwei Feng[1], Meng Liu[1], Bei Wang[1], Jun Feng[1,*], Jian Han[2], Jianni Liu[2]

* Correspondence: Jun Feng[1]

Email address: fengjun@nwu.edu.cn

Author details

[1] Department of Information Science and Technology, Northwest University, Xi'an 710127, China; hwfeng@nwu.edu.cn (H.F.); liumeng@stumail.nwu.edu.cn (M.L.);wangbei@stumail.nwu.edu.cn (B.W.);fengjun@nwu.edu.cn (J.F.)

[2] Early Life Institute, State Key Laboratory of Continental Dynamics, Department of Geology,

Northwest University, Xi'an 710069, China;

elihanj@nwu.edu.cn (J.H.); eliljn@nwu.edu.cn (J.L.)

References

Aberer AJ, Krompass D, Stamatakis A. 2013. Pruning Rogue Taxa Improves Phylogenetic Accuracy: An Efficient Algorithm and Webservice. *Systematic Biology* 62:162.

Baum BR, Estabrook GF. 1996. Impact of Outgroup Inclusion on Estimates by Parsimony of Undirected Branching of Ingroup Phylogenetic Lines. *Taxon* 45:243–257.

Bouamer S, Morand S. 2003. Phylogeny of Palaearctic Pharyngodonidae parasite species of Testudinidae: a morphological approach. *Canadian Journal of Zoology* 81:1885–1893.

Chen Y, Lin X, Huang M, et al. 2015. A new species of Lamiogethes and a new species of Meligethes from China (Coleoptera: Nitidulidae: Meligethinae). *Zootaxa* 3999:413.

Congreve CR, Lamsdell JC. 2016. Implied weighting and its utility in palaeontological datasets: a study using modelled phylogenetic matrices. *Palaeontology* 59:447–462.

Desper R, Gascuel O. 2002. Fast and accurate phylogeny reconstruction algorithms based on the minimum-evolution principle. *Journal of Computational Biology A Journal of Computational Molecular Cell Biology* 9:687–705.

Farabee MJ. 2007. Construction of Phylogenetic Trees. *Science* 155:279–284.

Farris JS. 2010. The Retention Index And The Rescaled Consistency Index. *Cladistics-the International Journal of the Willi Hennig Society* 5:417–419.

Felsenstein. 1988. Phylogenies from molecular sequences: inference and reliability. *Annual Review of Genetics* 22:521–565.

Foote M, Miller AI, Raup DM, et al. 2007. *Principles of paleontology*. Macmillan.

Goloboff PA, Farris JS, Nixon KC. 2010. TNT, a free program for phylogenetic analysis. *Cladistics* 24:774–786.

Goloboff PA, Pol D. 2005. Parsimony and Bayesian phylogenetics. *Parsimony Phylogeny & Genomics*:148–161.

Goloboff PA, Torres A, Arias JS. 2017. Weighted parsimony outperforms other methods of phylogenetic inference under models appropriate for morphology. *Cladistics-the International Journal of the Willi Hennig Society*.

Han J, Morris SC, Ou Q, et al. 2017. Meiofaunal deuterostomes from the basal Cambrian of Shaanxi (China). *Nature* 542:228–231.

Huang DW. 1996. *An Introduction to Cladistics*. China Agriculture Press.

Lee DC, Bryant HN. 1999. A Reconsideration of the Coding of Inapplicable Characters: Assumptions and Problems. *Cladistics-the International Journal of the Willi Hennig Society* 15:373–378.

Lewis PO. 2001. A Likelihood Approach to Estimating Phylogeny from Discrete Morphological Character Data. *Systematic Biology* 50:913–925.

Lin X. 2015. Preliminary studies of the taxonomy andphylogeny of the genus meligethes (coleoptera:nitidulidae: meligethinae) from china. Northwest A&F University.

Liu J, Steiner M, Dunlop JA, et al. 2011. An armoured Cambrian lobopodian from China with arthropod-like appendages. *Nature* 470:526–30.

Liu J, Shu D, Han J, et al. 2010. Morpho-anatomy of the lobopod Magadictyon cf. haikouensis from the Early Cambrian Chengjiang Lagerstätte, South China. *Acta Zoologica* 88:279–288.

Maddison WP. 1993. Missing Data versus Missing Characters in Phylogenetic Analysis. *Systematic Biology* 42:576–581.

Mahapatra S, Sahu SS, Priyam A, et al. 2018. Predicting protein-RNA interaction using sequence derived features and machine learning approach. *International Journal of Data Mining & Bioinformatics* 19:270.

Nelson G., Platnick NI. 2010. Three-Taxon Statements: A More Precise Use of Parsimony? *Cladistics-the International Journal of the Willi Hennig Society* 7:351–366.

O'Reilly Joseph E, Puttick Mark N, Parry Luke, et al, Pisani Davide, Donoghue Philip C. J. 2016. Bayesian methods outperform parsimony but at the expense of precision in the estimation of phylogeny from discrete morphological data: *Biology Letters* 12:20160081.

O'Reilly Joseph E, Puttick Mark N, Pisani D, et al. 2018. Probabilistic methods surpass parsimony when assessing clade support in phylogenetic analyses of discrete morphological data. *Palaeontology* 61:105–118.

Platnick NI., Griswold CE., Coddington JA. 2010. On Missing Entries In Cladistic Analysis. *Cladistics-the International Journal of the Willi Hennig Society* 7:337–343.

Puttick MN, O'Reilly JE, Tanner AR, et al. 2017. Uncertain-tree: discriminating among competing approaches to the phylogenetic analysis of phenotype data. *Proceedings Biological Sciences* 284.

Robinson DF, Foulds LR. 1981. Comparison of phylogenetic trees. *Mathematical Biosciences* 53:131–147.

Saitou N, Nei M. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution* 4:406.

Scott R, Gras R. 2012. Comparing Distance-Based Phylogenetic Tree Construction Methods Using An Individual-Based Ecosystem Simulation, EcoSim. In: *Artificial Life*.

Seitz V, Garcia SO, Liston A. 2000. Alternative Coding Strategies and the Inapplicable Data Coding Problem. *Taxon* 49:47–54.

Sneath PHA, Sokal RR. 1973. *Numerical Taxonomy: The Principles and Practice of Numerical Classification*.

Strong EE, Lipscomb D. 2010. Character Coding and Inapplicable Data. *Cladistics-the International Journal of the Willi Hennig Society* 15:363–371.

Tamura K, Nei M, Kumar S. 2004. Prospects for inferring very large phylogenies by using the neighbor-joining method. *Proc.natl.acad.sci.usa* 101:11030–11035.

Tang LD, Yuan MM, Li Y, et al.. Phylogenetic Analysis of Hibiscus Based on Morphological Characters. *Journal of Henan Agricultural Sciences* 43:105–111.

Tian Y, Kubatko L. 2017. Rooting phylogenetic trees under the coalescent model using site pattern probabilities: *Bmc Evolutionary Biology* 17:263.

Wilkinson M. 2010. A comparison of two methods of character construction. *Cladistics-the International Journal of the Willi Hennig Society* 11:297–308.

Yang J, Ortegahernández J, Gerber S, et al. 2015. A superarmored lobopodian from the Cambrian of China and early disparity in the evolution of Onychophora. *Proc Natl Acad Sci U S A* 112:8678–8683.

Yang Z, Rannala B. 2012. Molecular phylogenetics: principles and practice. *Nature Reviews Genetics* 13:303.

Zaragüetabagils R, Bourdon E. 2007. Three-item analysis: Hierarchical representation and treatment of missing and inapplicable data. *Comptes rendus - Palevol* 6:527–534.

Figures

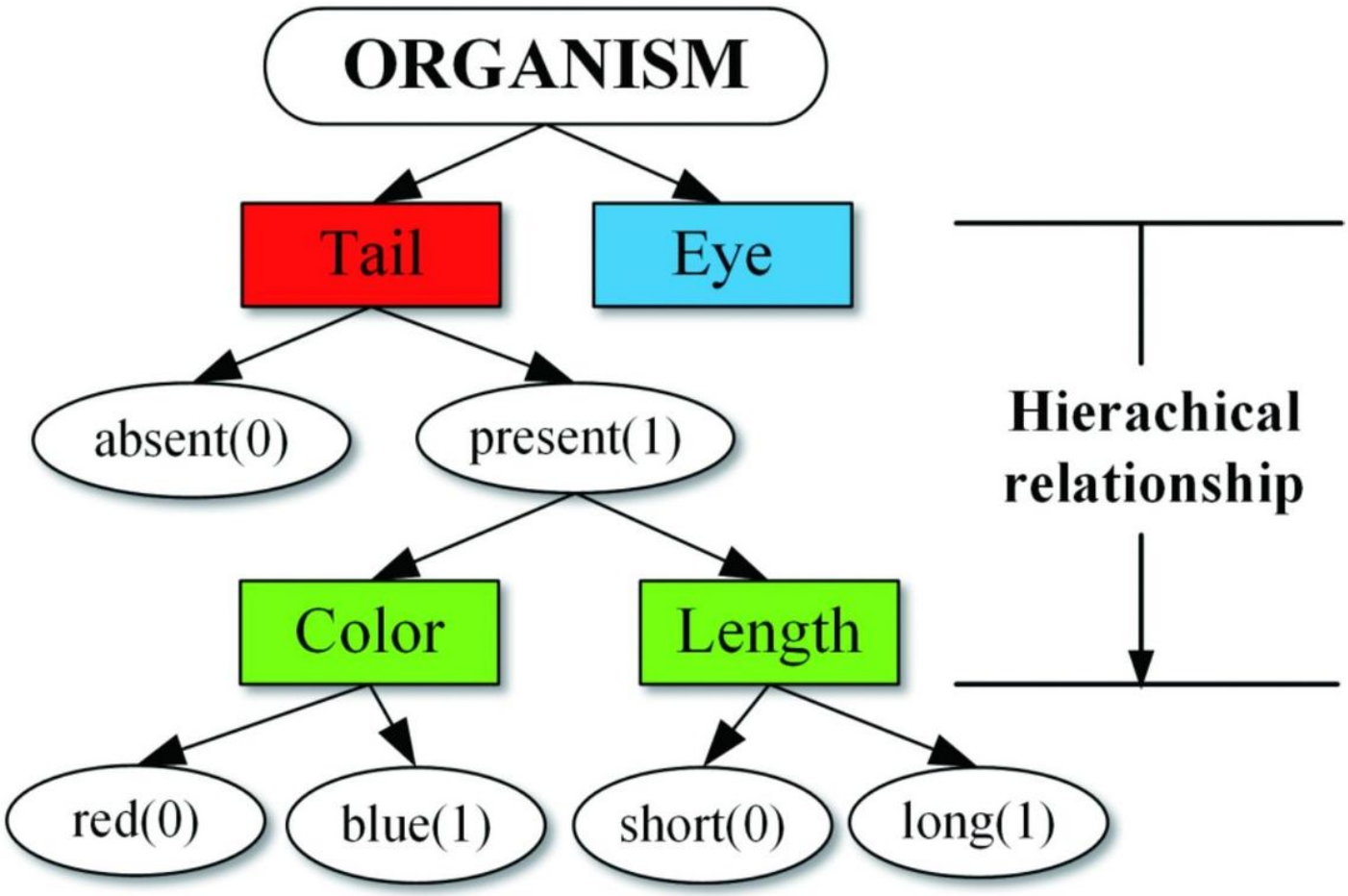


Figure 1

Example of the hierarchical relationship among characters. The rectangles represent the characters and the ellipses represent the states of the character. The red, green and blue rectangles represent the upper character, lower character, and solitary character, respectively.

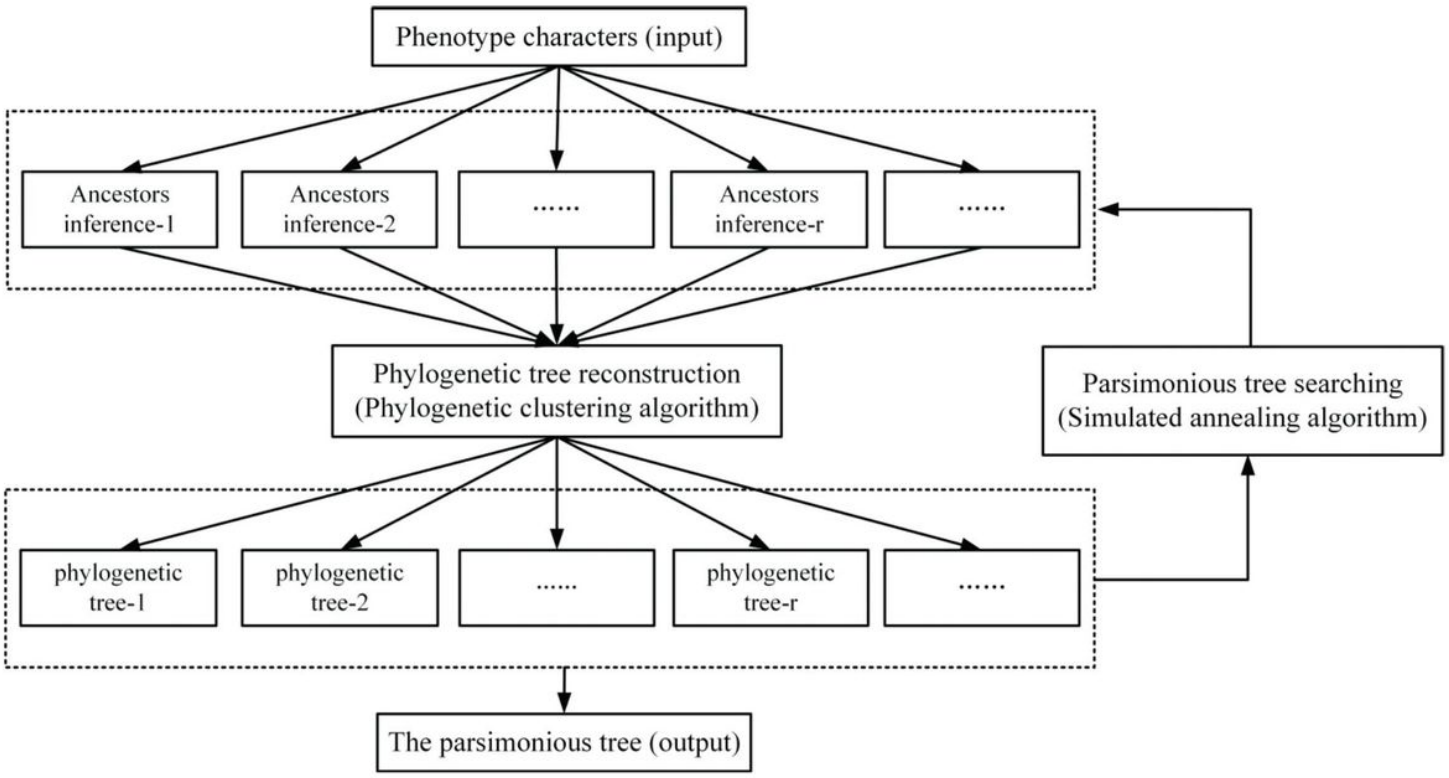


Figure 2

The framework of phylogenetic analysis method based on HCPC.

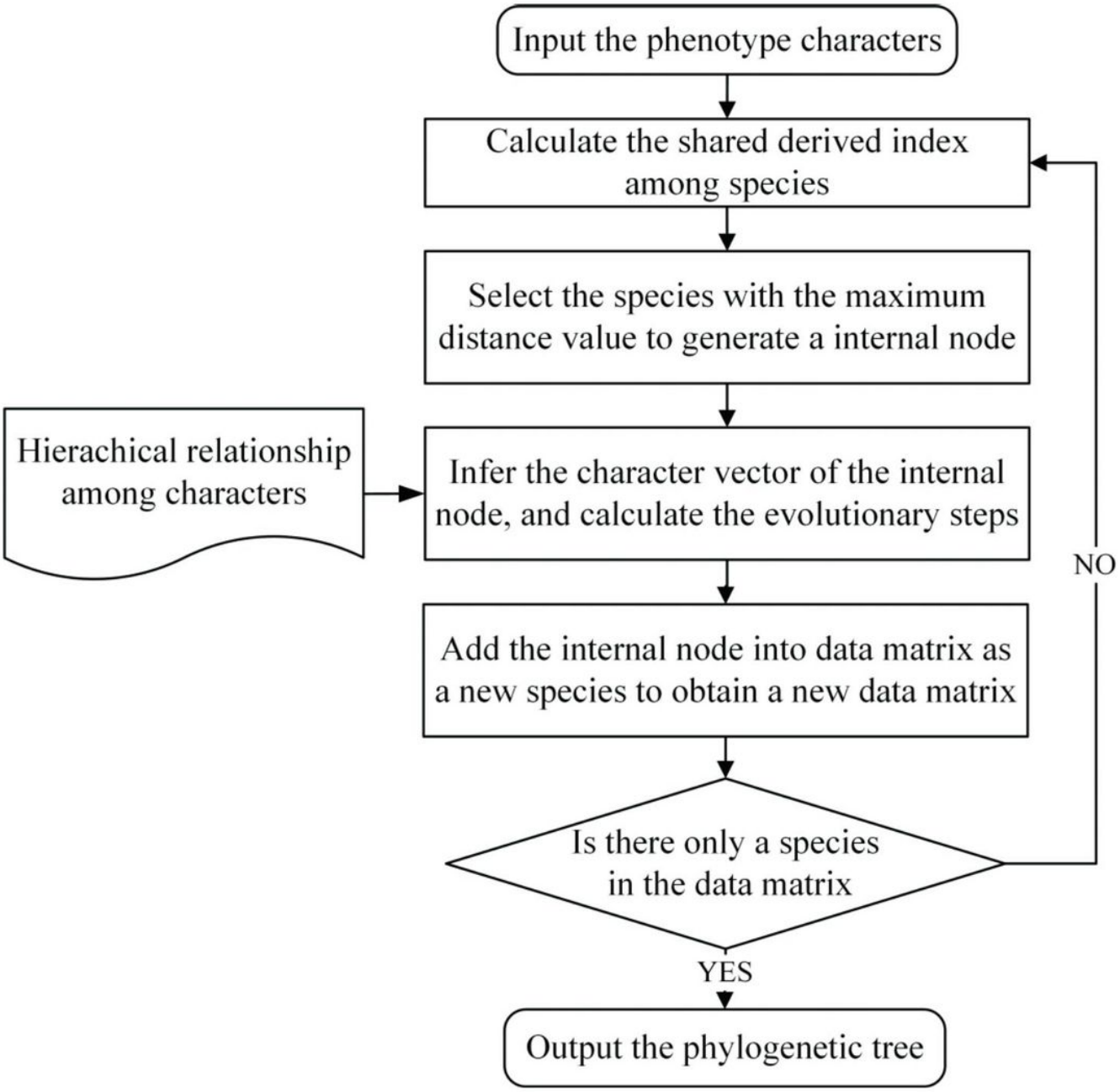


Figure 3

Flowchart of the phylogenetic clustering based on the character hierarchy for the phylogenetic reconstruction.

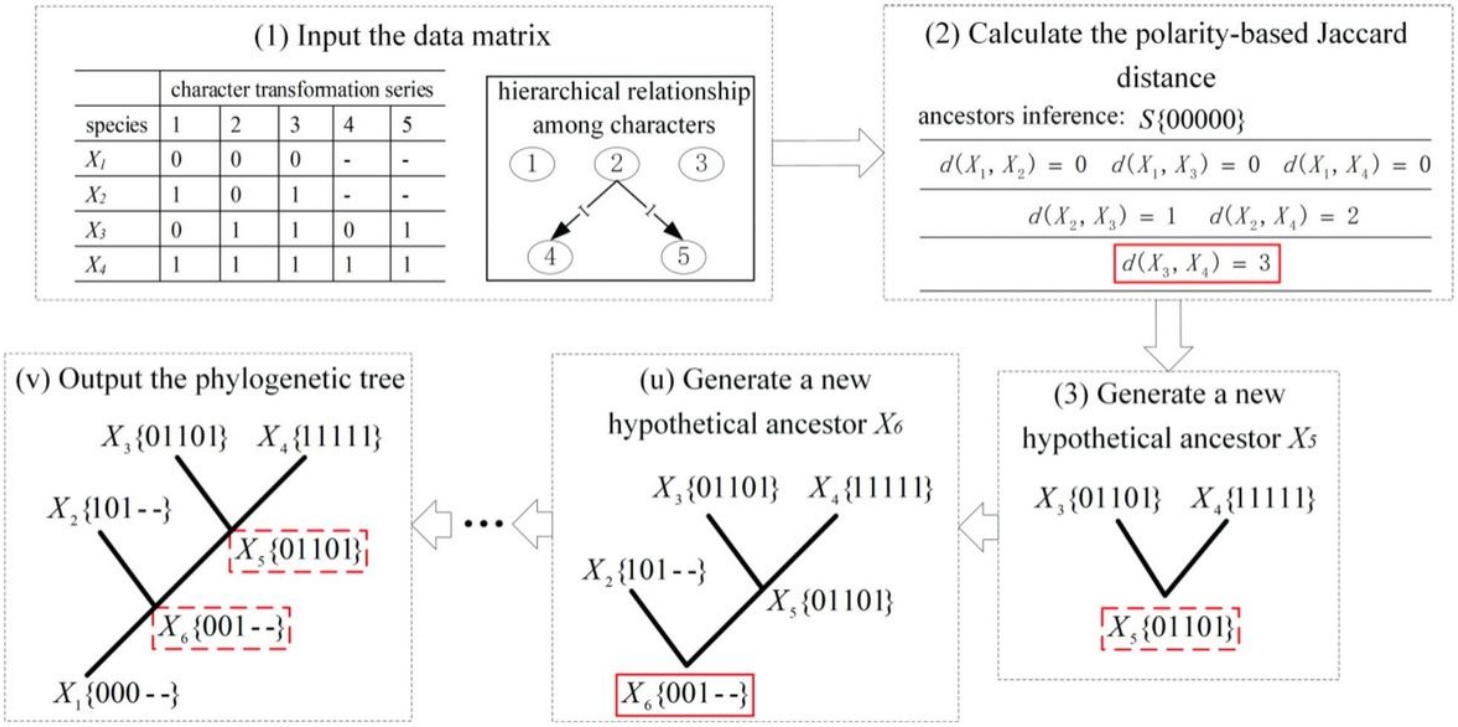


Figure 4

Diagram of an example of Phylogenetic Reconstruction. The red solid line box is the maximum shared derived index and the red dotted boxes are the hypothetical species.

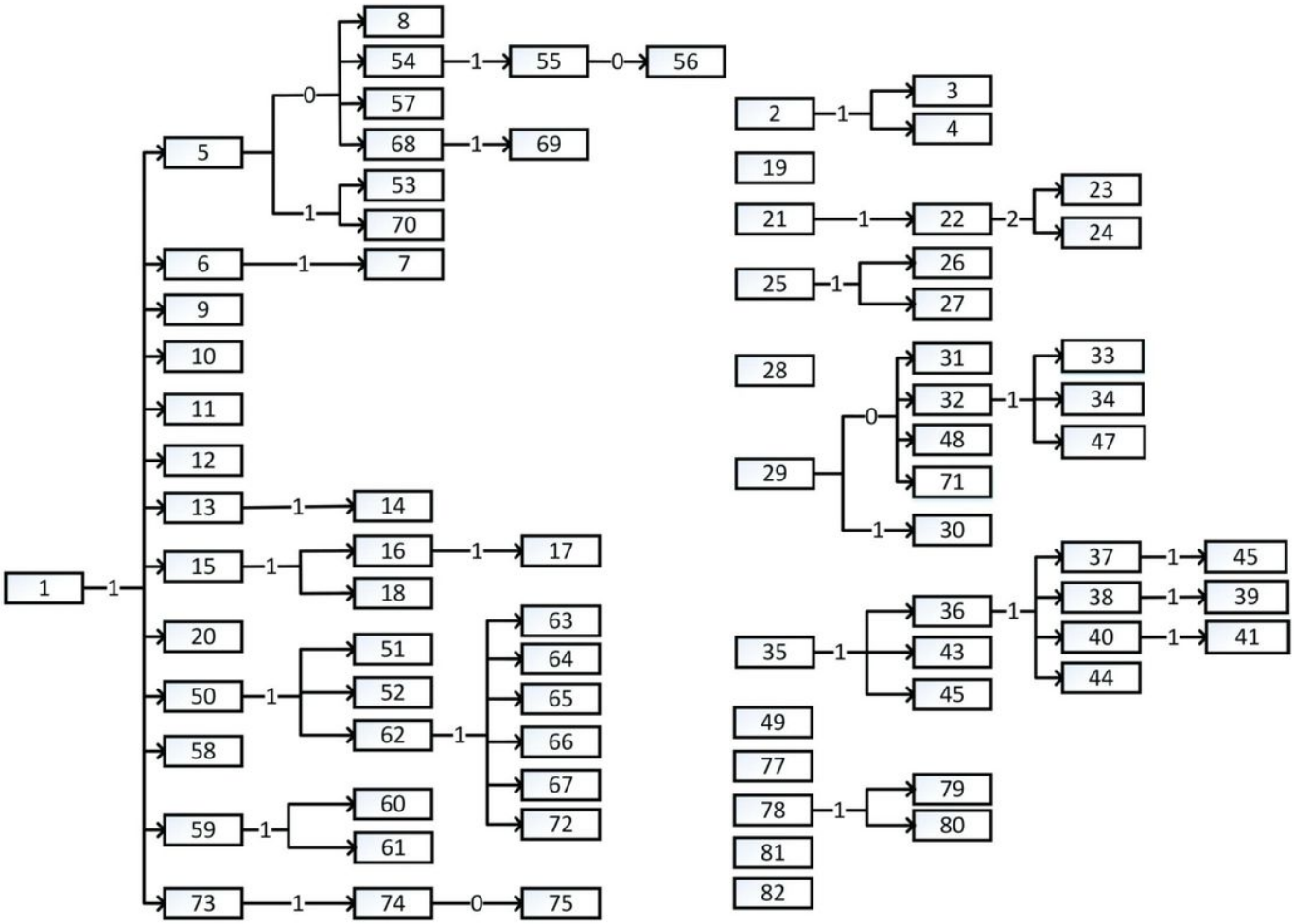


Figure 5

A diagram of the hierarchical structure of characters for paleozoic lobopodians (part). The number in the box represents the No. character and the number on the arrow represents the character state.

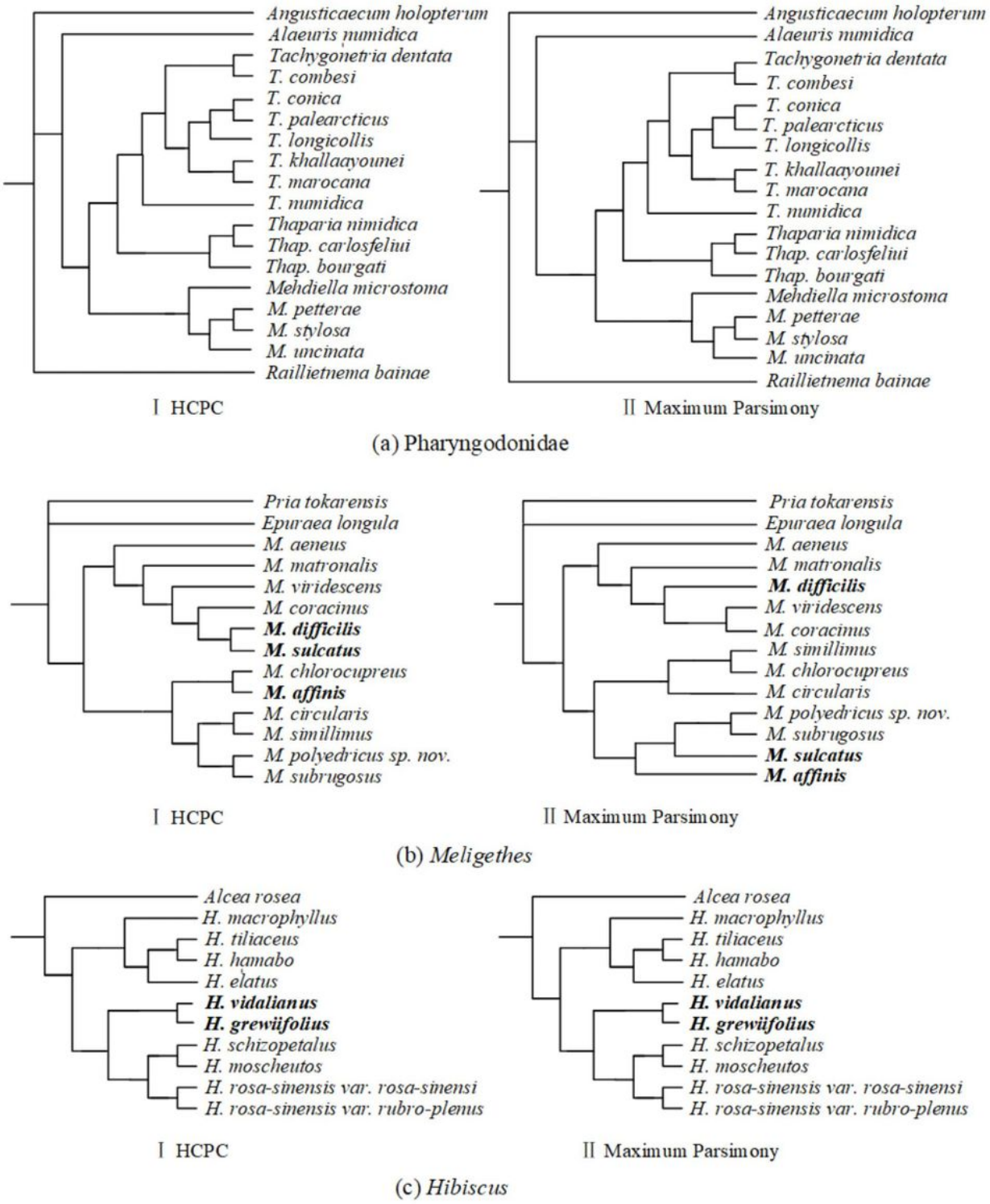


Figure 6

A comparison of the phylogenetic trees reconstructed by HCPC and MP for living species. (A), (C), and (E) are the phylogenetic trees constructed using the HCPC method for the Bouamer & Morand dataset (Pharyngodonidae), the Chen et al. dataset (Meligethes), and the Tang et al. dataset (Hibiscus), respectively. (B), (D), and (F) are the phylogenetic trees constructed using the MP method for the Bouamer & Morand dataset (Pharyngodonidae), the Chen et al. dataset (Meligethes), and the Tang et al. dataset

(Hibiscus), respectively. Species with different positions on the inferred trees and the model trees are marked in bold.

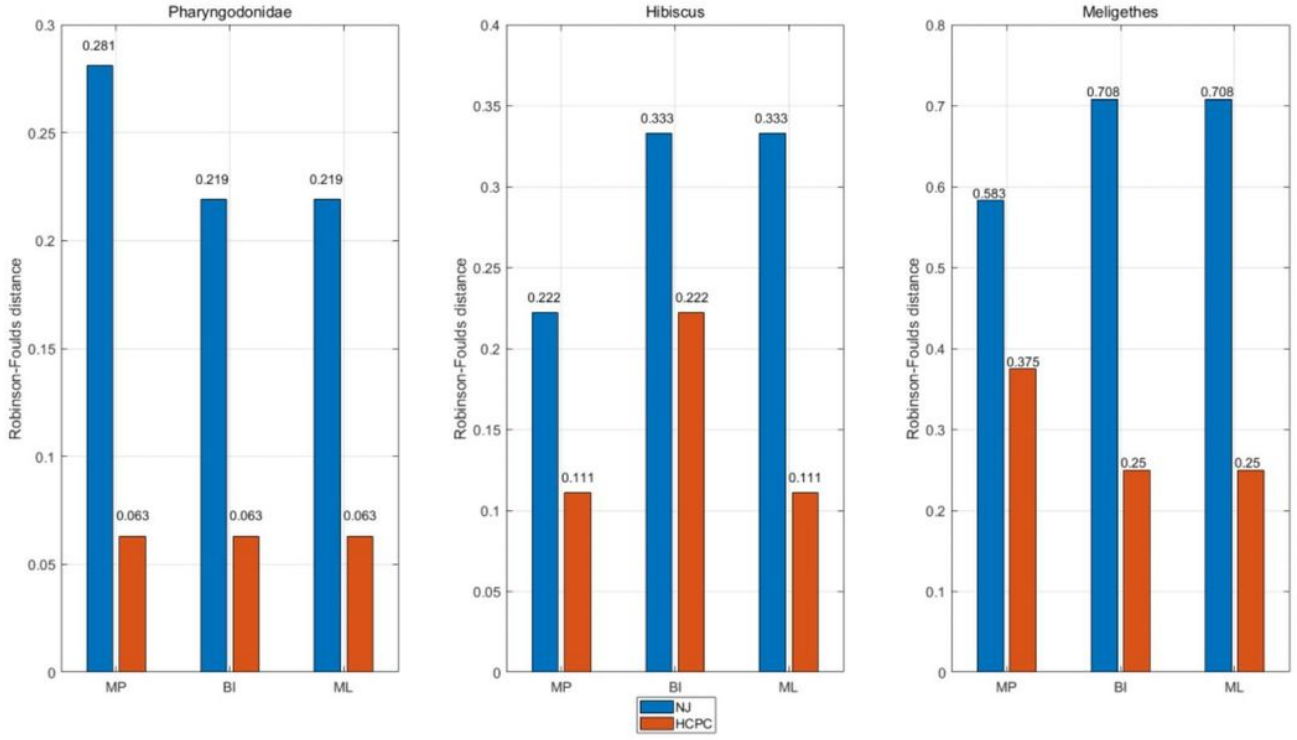


Figure 7

RF distance for the NJ method and HCPC for living species.

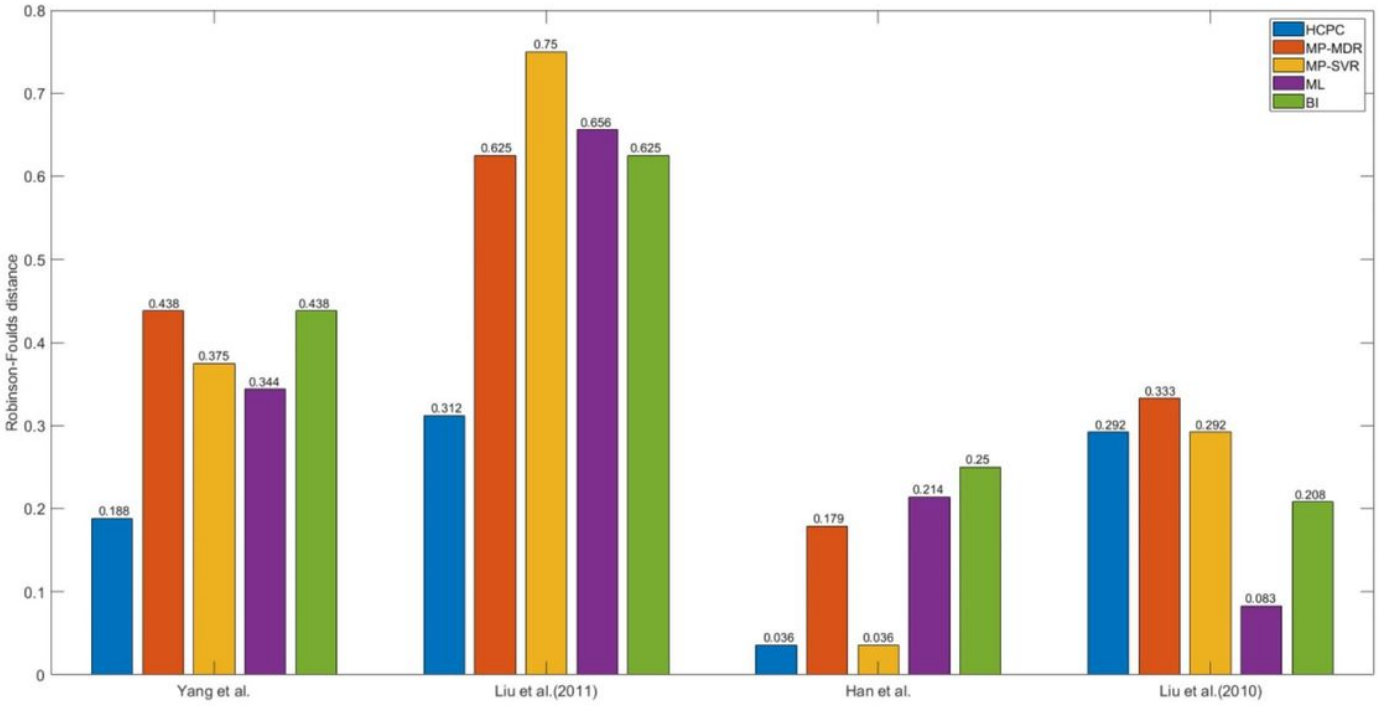


Figure 8

RF distance for the HCPC and other methods for paleontological species. The MP-MDR, ML, and BI methods use the MDR to treat inapplicable states; the MP-SVR uses SVR to treat inapplicable states.