CHAPTER 11

ANALYSIS OF PHYLOGENY: A CASE STUDY ON SAURURACEAE

Shao-Wu Meng

Institute for Infocomm Research
swmeng@i2r.a-star.edu.sg

Phylogenetics is the study of the origin, development, and death of a taxon. It is a useful tool, for example, in the conservation of species. This chapter is an introduction to phylogenetics from the perspective of plant molecular biologists, using a case study on Saururaceae. Based on analysis using integrated data from DNA sequences and morphology, we also draw a surprising conclusion that *Saururus* is not the most primitive genus in Saururaceae, contradicting a long-held tradition of the field. We point out some deficiencies in the older studies.

ORGANIZATION

Section 1. We begin with a brief explanation of what is phylogeny, why do people study phylogeny, and how do people study phylogeny.

Section 2. To illustrate the description of phylogenetics in Section 1, we do a case study on the phylogeny of the relic paleoherb Saururaceae. A background of Saururaceae and the case study is given.

Section 3. A summary of the materials and methods used in the case study is presented. The methods include the steps of collecting plant materials, extracting and sequencing DNA from these materials, aligning these DNA sequences, and performing parsimony analysis of these sequences and morphological data.

Section 4. Then we present the results of each step in detail, focusing especially on the phylogentic trees constructed by parsimony analysis on data from 18S nuclear genes, from *trn*L-F chloroplast DNA sequences, from *mat*R mitochondrial genes, from a combined matrix of these genes, and from morphological data.

Section 5. After that, we dive into an extended discussion on the phylogeny of Saururaceae. On the basis of our analysis, we conclude that Saururacease, *Saururus*, and *Gymnotheca* are monophyly, that *Anemopsis* and *Houttuynia* are sister group, that *Saururus* and *Gymnotheca* are also sister group, and that the *Anemopsis-Houttuynia* form the first clade of Saururaceae. This is in disagreement with several earlier studies that *Anemopsis* and *Houttuynia* are derived from *Saururus*,[631] that *Saururus* is the first to diverged from the ancestral Saururaceous stock,[846] *etc.* We point out some of the deficiencies of these earlier studies.

Section 6. Finally, we provide some advice on reconstructing phylogeny, spanning the aspects of sampling of species, selection of out-group, alignment of sequences, choosing phylogeny reconstruction methods and parameters, dealing with morphological data, and comparing phylogenies from molecular data and morphological data.

1. The What, Why, and How of Phylogeny

1.1. *What is Phylogeny?*

Phylogeny is the evolutionary process of an organism from its initial occurrence to a concrete geological time. It includes the evolutionary process, and also the organism itself and its descendants. Phylogenetics is a research field to study the phylogeny of organisms. It is impossible for a researcher to study the phylogeny of all extinct and existing organisms. Phylogenetists generally study only the origin, development, and death of one taxon.

1.2. *Why Study Phylogeny?*

The aim of studying phylogeny is to reconstruct the evolutionary process and the phylogenetic relationship among the descendants of an organism. It is useful as described in the following two points.

First, the study on phylogeny can satisfy the curiosity of the scientist and the public. *E.g.*, which species is the closest relative of Man, and when did Man separate from it? After careful phylogenetic analysis of extensive data, including DNA-DNA hybridization[129] and mitochondrial DNA sequence,[364] the closest extant relatives of Man were identified as two Chimpanzee species, followed by gorillas, orang utans, and the nine gibbon species.[501] According to microcomplement fixation data, Sarich and Wilson[744] estimated that the divergence time between Human and Chimpanzee was 5 million years ago, rather than 15 million years ago, which was commonly accepted by paleontologists at that time. Molecular phylogenetics has also been greatly pushed forward by answering these questions.

Second, the study on phylogeny can guide us today. *E.g.*, the dusky seaside sparrow, scientifically named *Ammodramus maritimus nigrecens*, had habited the salt marshes in Brevard County, Florida. By 1980, there were only six individuals, and all of them were male. In order to conserve the subspecies, an artificial breeding program was launched. First of all, the key of the program was to find out the phylogenetically closest subspecies to *A. M. nigrecens*. The subsequent steps was to mate females from their closest subspecies with males from *A. M. nigrecens*, then to mate the female hybrids of the first generation with the males of *A.*

M. nigrecens, and then to mate the female hybrids of the second generation with the males of *A. M. nigrecens*, and so on, as long as the males of *A. M. nigrecens* lived.[501] All effort would be useless if the closest subspecies was chosen wrongly. Hence, phylogenetic analysis played an important role in such a program.

In short, phylogeny study is a useful tool, not only in theoretical study, but also in practical use. It is very helpful for a study that needs to know evolutionary history or the phylogenetic relationship among organisms.

1.3. *How to Study Phylogeny*

According to the definition of phylogeny, most contents of phylogeny are invisible because they have been historical. Fortunately, a few extinct organisms have been fossilized and kept in different stratums due to physical or chemical reaction of geology. Moreover, a part of these fossils have been dug out from different stratums by paleontologists. The visible contents of phylogeny—the basis upon which to reconstruct the evolutionary process—are the existing organisms now and those fossils that have been dug out.

Paleontologists reconstruct phylogeny according to fossils, DNA sequences in fossils, and geological accidents.[486] The reconstruction process and its conclusions can be reliable if paleontologists have enough fossils. It is unfortunate that fossils are often rare and not full-scale when compared with the extinct organisms in geological time. Anyway, fossils are very important to reconstruct a time frame of evolution.

Other phylogenetists reconstruct phylogeny mainly according to the characters of extant organisms. Since characters of extant organisms are the results of evolution, they should reflect evolutionary history. Morphological phylogenetists reconstruct phylogeny mainly according to morphological characters from gross morphology, anatomy, embryology, cytology, physiology, chemistry, *etc*.[545,846] Molecular phylogenetists reconstruct phylogeny mainly according to isozyme, DNA sequences, protein sequences, *etc*.[239,688,786]

Recently, phylogenetic reconstruction by comparing sequences of whole genomes gradually becomes fashionable. The advantage of comparing whole genomes is that it can avoid unilateral results from only one or a few types of DNA sequences.[942] Even though this method is important in molecular phylogenetics, comparing whole genomes should not be overwhelming. For example, the sequence difference in genomes between Human and Chimpanzee is less 0.2%. However, the phenotype difference between them is great. What is the reason? It is possible that the DNA sequence difference of less 0.2% is fatal, or there are other mechanisms, such as differences in secondary structure of RNAs and proteins, to decide the phenotype difference between Human and Chimpanzee. Moreover, phenotype is affected not only by genotype, but also by environment. Since a conclusion drawn only from DNA sequence comparison ignores the environmental differences, it could be biased.

Each of the ways above has its own advantages and disadvantages. Hence, they can complement each other. So a better way is to study phylogeny based on integrated data, that come from different subjects and from different levels. In subsequent sections, we present a case study of the phylogeny of Saururaceae on the basis of integrated data from three types of DNA sequences from three genomes and 58 morphological characters.

2. Case Study on Phylogeny of Saururaceae

Saururaceae is a core member of the paleoherbs.[845] It is an ancient and relic family with six species in four genera, *viz. Saururus*, *Gymnotheca*, *Anemopsis*, and *Houttuynia*.[505] They are perennial herbs with simple flowers that bear bracts without perianths. Saururaceae is an East Asian-North American disjunctive family, with *Anemopsis* and *Saururus cernuus* in North America, and *Houttuynia*, *Gymnotheca*, and *Saururus chinensis* in East Asia. Due to its important systematic position and interesting geographical pattern of distribution, Saururaceae has been a hot spot for phylogenetists even though it is a small family having just a few species.

The viewpoints on the phylogeny of Saururaceae are very different based on morphology, including gross morphology, cytology, floral morphogensis, *etc.* Wu and Wang[908] included *Saururus*, *Circaeocarpus*, *Anemopsis*, *Houttuynia*, and *Gymnotheca* in Saururaceae. They thought that *Circaeocarpus* was derived from *Saururus* firstly, *Anemopsis* secondly, *Gymnotheca* thirdly, and *Houttuynia* fourthly. Later, they[909] detected that the newly published genus, *Circaeocarpus*, was in fact a member of Piperaceae, and *Circaeocarpus saururoides* C. Y. Wu and *Zippelia begoniaefolia* Blume were conspecific. From the point of view of plant biogeography, Wu[906] later thought *Anemopsis* and *Houttuynia* were vicariant genera, and *S. chinensis* and *S. cernuus* were vicariant species. Based on the basic chromosome numbers of some genera in Saururaceae, Okada[631] put forward that *Anemopsis* and *Houttuynia* were respectively derived from *Saururus*, and they were at the same advanced level. Lei *et al.*[484] supported Okada's opinion, and thought that *Gymnotheca* was the most advanced genus. On the basis of a cladistic analysis of morphological and ontogenetic characters, Tucker *et al.*[846] made an estimate that *Saururus* was the first to diverge from the ancestral Saururaceous stock. They also suggested this was followed by *Gymnotheca*, with *Hout-*

tuynia and *Anemopsis* being sister taxa. Combining the data from gross morphology, anatomy, embryology, palynology, cytology, and flower development, Liang[505] proposed that the ancestor of Saururaceae was divided into two branches at early times. One was the *Gymnotheca-Anemopsis*. The other was the *Saururus-Houttuynia*. Some genera of Saururaceae have been represented in recent studies on molecular phylogeny of higher-level within angiosperm.[139, 688, 786, 787] However, they have not studied the phylogeny of Saururaceae.

In short, although several scientists have done a lot of research on Saururaceae, there is still no uniform opinion on the phylogeny of Saururaceae. Hence, we try to construct a more reliable phylogeny of Saururaceae. Our study is based on three types of DNA sequences from all three genomes and 58 stable morphological characters. The three types of DNA sequences are: 18S functional gene from nuclear genome, *trn*L-F DNA sequence from chloroplast genome, and *mat*R functional gene from mitochondrial genome.

18S and *mat*R are generally used for reconstructing higher-level phylogeny, such as relationships of orders, families, or distant genera.[688, 787] *Trn*L-F are commonly used for genera, species, and lower levels. For studying the phylogenetic relationships within Saururaceae, an ancient and relic family, we select the three types of DNA sequences. Meanwhile, we select 58 morphological characters to rebuild the phylogeny of Saururaceae, and to compare with the phylogenies of previous studies on Saururaceae. These morphological characters are stable and come from gross morphology, anatomy, embryology, palynology, cytology, and flower development.

3. Materials and Methods

3.1. *Plant Materials*

We collect from natural populations or cultivated plants all six species of the in-group, *Anemopsis californica*, *Gymnotheca chinensis*, *Gymnotheca involucrata*, *Houttuynia cordata*, *S. cernuus*, and *S. chinensis*; and three designated out-groups, *Peperomia tetraphylla*, *Piper mullesua*, and *Z. begoniaefolia* (all Piperaceae). Then, we deposit vouchers in the herbarium of the Kunming Institute of Botany, Chinese Academy of Sciences, Kunming, Yunnan Province, People's Republic of China.

3.2. *DNA Extraction, PCR, and Sequencing*

A detailed description of the DNA extraction, PCR, and sequencing steps are given in Meng *et al.*[560]

3.3. *Alignment of Sequences*

We check each DNA sequence with its electrophoretic map using the SeqEd program (Applied Biosystems), and decide the ends of 18S gene by comparing with AF206929 (from GenBank, the same thereafter), the ends of *trn*L-F by comparing with AF200937, and the ends of *mat*R gene by comparing with AF197747, AF197748 and AF197749. Using the Clustal-X program[825] and MEGA2b3,[464] we align all our sequences.

3.4. *Parsimony Analysis of Separate DNA Sequences*

Parsimony is one of several criteria that may be optimised in building phylogenetic tree. The key idea of parsimony analysis is that some trees fit the character-state data better than other trees. Fit is measured by the number of evolutionary character-state changes implied by the tree. The fewer changes the better.

We analyze the aligned sequences using maximum parsimony in PAUP.[810] More concretely, we use branch-and-bound search with random addition sequence and ACCTRAN character state optimization, and treat gaps as missing data. The number of replicates in bootstrap analysis is 1000.

3.5. *Parsimony Analysis of Combined DNA Sequences*

The mutation in 18S gene is so slow that the 18S sequences between *Peperomia tetraphylla* and *Peperomia serpens* are almost identical. Hence, we replace *Pe. serpens* by *Pe. tetraphylla* in the alignment of 18S gene. Ditto for *mat*R genes. Therefore, we combine all sequence data into one alignment. Using maximum parsimony analysis and branch-and-bound search, we do a partition-homogeneity test for different parts of the combined data. The partition-homogeneity test checks the homogeneity among different parts of a matrix; it is useful for analyzing a data matrix that is combined from different data. When executing the test, the number of replicates is set to 1000. Then, we analyze the combined data by using the same settings as when analyzing separate DNA alignments. Alignments of each gene or combined DNA sequences are available upon request from us.

3.6. *Parsimony Analysis of Morphological Data*

We selected 58 morphological characters to reconstruct the phylogeny of Saururaceae. These characters are given in Figures 7–9. The characters are from studies of herbarium specimens and literature.[484, 502–507, 561, 839–844, 846] These characters pertain to gross morphology, anatomy, embryology, palynology, cytology, and flower development. Concretely, 2 characters are from cytology, 11 characters are from vegetative organs, and 45 characters are from reproductive organs. Moreover, we treat 34 of these characters as binary characters and 24 of these characters as multi-state characters, as detailed in Figure 10. We designate *Z. begoniaefolia, Piper*, and *Peperomia* as out-group. As before, we used PAUP[810] to analyze the morphological matrix of Saururaceae. All characters are un-weighted and unordered. Other settings are the same as when analyzing the DNA sequences.

3.7. *Analysis of Each Morphological Characters*

Using WINCLADA,[619] we analyzed each morphological character in order to know which one is homologous and which one is homoplasious. We used maximum parsimony analysis and the following setting: heuristics search, 1000 replications, 1 starting tree per replication, multiple TBR and TBR search strategy, 0 random seed, and slow optimization.

Here, TBR is an acronym for tree-bisection-reconnection. It is a heuristic algorithm for searching through treespace. It proceeds by breaking a phylogenetic tree into two parts and then reconnecting the two subtrees at all possible branches. If a better tree is found, it is retained and another round of TBR is initiated. This is quite a rigorous method of searching treespace.

4. Results

4.1. *Phylogeny of Saururaceae from 18S Nuclear Genes*

Alignment of 18S gene sequences produces a matrix of 1567 positions. 64 of these positions are variable-uninformative; that is, each of these 64 columns of the alignment has two or more types of bases, but at most one of these types of bases has two or more individuals. 35 of these positions are parsimony-informative; that is, each of these 35 columns of the alignment has two or more types of bases with at least two individuals each. The remaining 1468 positions are constant and uninformative. The percentage of parsimony-informative sites, calculated as the ratio of the number of parsimony-informative positions to the total number of positions, is 2.23% ($= 35/1567$).

Our maximum parsimony analysis produces two most parsimonious trees of 123 steps. The strict consensus of the two trees is depicted in Figure 1. Saururaceae (98%) (bootstrap value, the same thereafter), *Gymnotheca* (99%) and *Saururus* (70%) are monophyly; that is, each of these groups have an immediate common ancestor. *A. californica* is the sister group of *H. cordata* (75%), and they formed the basal of Saururaceae. *Saururus* is the sister group of *Gymnotheca* (99%).

Bootstrap is a method to estimate the confidence levels of inferred relationships. The process of bootstrapping is to iteratively create a new data matrix, that has the same size as the original data matrix, by randomly resampling individual columns (ı.e., characters) of the original data matrix. In this process, some characters of the original data matrix may be sampled more than once and some may not be sampled at all. This process is repeated many times and phylogenies are reconstructed each time. After all these processes of bootstrapping are finished, a majority-rule consensus tree is constructed from the optimal tree from each bootstrap process. The bootstrap support value—*i.e.*, the bootstrap value—for any internal branch is the number of times that it was recovered during these processes of bootstrapping. Generally, the bootstrap process should be repeated over 500 times. If the bootstrap value of a branch is above 70%, the branch can be regarded as a reliable one because, according to the simulation study of Hillis and Bull,[352] this bootstrap value corresponds to a probability of 95% that the branch is real.

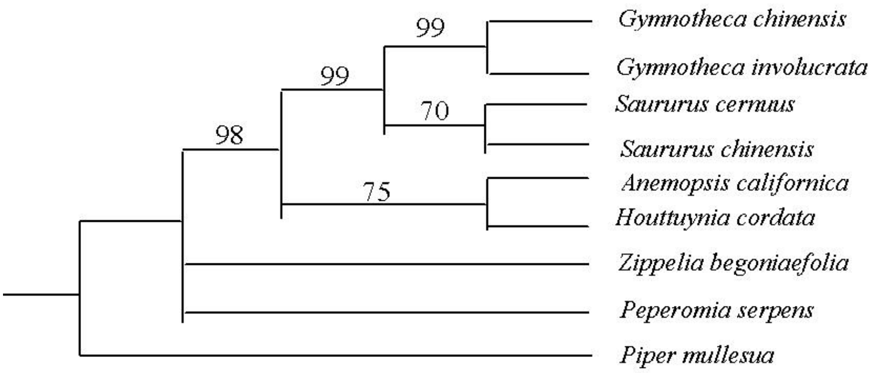

Fig. 1. The strict consensus of the two most parsimonious trees of Saururaceae based on 18S nuclear genes. Length = 123, CI =0.8943, RI = 0.7833, RC = 0.7005. Number of bootstrap replicates = 1000. Bootstrap values (%) are above branches. The Consistency Index (CI) is a measure of how well an individual character fits on a phylogenetic tree. It is calculated by dividing the minimum possible number of steps by the observed number of steps. If the minimum number of steps is the same as the observed number of steps, then the character has a CI of 1.0 (perfect fit). If a character is not completely compatible with a tree, then it has a CI value less than 1.0 or even approaching zero (poor fit). The CI value of a tree is the average CI value over all of the characters. There is a problem with this value: It is always 1.0 for autapomorphies, which are character states that are seen in a single sequence and no other. The Retention Index (RI) is similar to CI, but is also more resistant to bias due to autopomorphies. The Re-scaled Consistency Index (RC) is also used to assess the congruency and fit of characters to a tree (range from 0 to 1). It is computed as $RC = CI * RI$. A higher value of RC indicates that the characters in the data set are more congruent with each other and with the given tree.

4.2. *Phylogeny of Saururaceae from trnL-F Chloroplast DNA Sequences*

Alignment of *trn*L-F DNA sequences produces a matrix of 1198 positions. 81 of these positions are variable and uninformative. 92 of these positions are parsimonious and informative. The percentage of parsimony-informative sites is 7.679%.

Our maximum parsimony analysis produces the single most parsimonious tree of 203 steps depicted in Figure 2. Saururaceae (100%), *Gymnotheca* (100%), and *Saururus* (100%) are monophyly. *A. californica* is the sister group of *H. cordata* (100%), and they form the first clade of Saururaceae. *Saururus* is the sister group of *Gymnotheca* (71%).

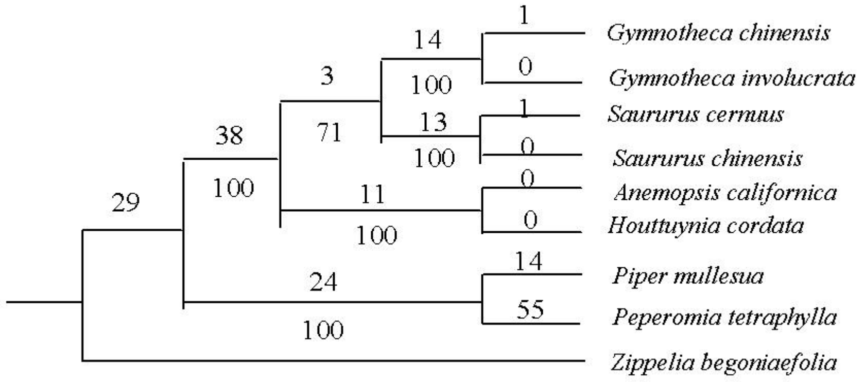

Fig. 2. The single most parsimony tree of Saururaceae based on *trn*L-F chloroplast DNA sequences. Length = 203, CI =0.9458, RI = 0.9231, RC = 0.8731. Number of bootstrap replicates = 1000. Base substitution values are shown above the branches, and bootstrap values (%) are shown below the branches.

4.3. *Phylogeny of Saururaceae from matR Mitochondrial Genes*

Alignment of *mat*R gene sequences produces a matrix of 1777 positions. 118 of these positions are variable. 43 of these positions are parsimonious and informative. The percentage of parsimony-informative sites is 2.42%.

Our maximum parsimony analysis yields the single most parsimonious tree of 136 steps depicted in Figure 3. Saururaceae (96%), *Gymnotheca* (100%), and *Saururus* (99%) are monophyly. *A. californica* is the sister group of *H. cordata* (83%), and they form the first clade of Saururaceae. *Saururus* is the sister group of *Gymnotheca* (95%).

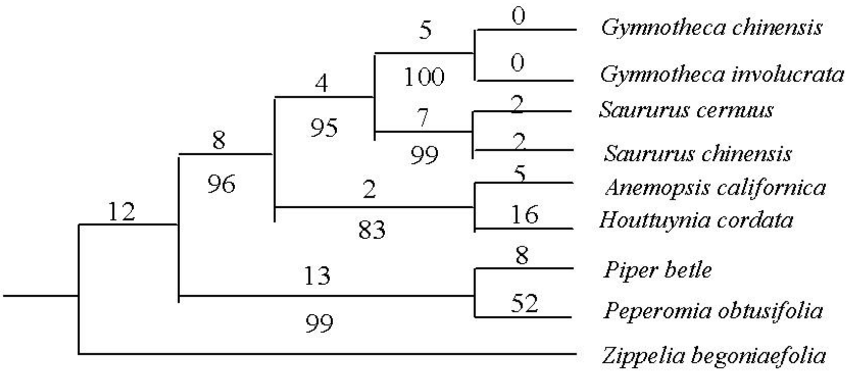

Fig. 3. The single parsimony tree of Saururaceae based on *mat*R mitochondrial genes. Length = 136, CI =0.9118, RI = 0.8125, RC = 0.7408. Number of bootstrap replicates = 1000. Base substitution values are shown above the branches, and bootstrap values (%) are shown below the branches.

4.4. *Phylogeny of Saururaceae from Combined DNA Sequences*

Alignment of all DNA sequences produces a matrix of 4542 positions. 199 of these positions are variable-uninformative. 171 of these postions are parsimony-informative. The percentage of parsimony-informative sites is 3.76%. The P value of the partition-homogeneity test is 1.000000.

The single most parsimonious tree of 435 steps depicted in Figure 4 is produced by our maximum parsimony analysis. Saururaceae (100%), *Gymnotheca* (100%), and *Saururus* (100%) are monophyly. *A. californica* is the sister group of *H. cordata* (100%), and they form the first clade of Saururaceae. *Saururus* is the sister group of *Gymnotheca* (100%).

4.5. *Phylogeny of Saururaceae from Morphological Data*

1 character is constant. 16 characters are variable-uninformative. 41 characters are parsimony-informative. The percentage of parsimony-informative characters is 70.69%.

Our maximum parsimony analysis yields the single most parsimonious tree of 97 steps shown in Figure 5. Saururaceae (100%), *Gymnotheca* (92%), and *Saururus* (100%) are monophyly. *A. californica* is the sister group of *H. cordata* (65%), and they formed the first clade of Saururaceae. *Saururus* is the sister group of *Gymnotheca* (54%).

However, the bootstrap support values for *Gymnotheca*-*Saururus* and for *Anemopsis*-*Houttuynia* are somewhat weak. We think there is interference from

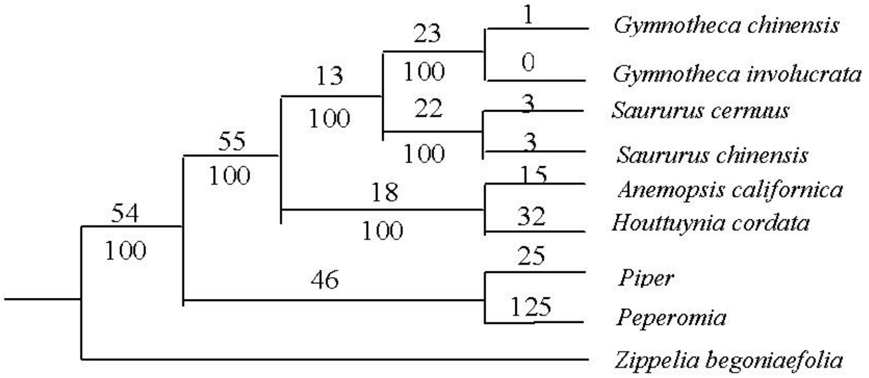

Fig. 4. The single parsimony tree of Saururaceae based on combined DNA sequences. Length = 435, CI =0.9264, RI = 0.8810, RC = 0.8162. Number of bootstrap replicates = 1000. Base substitution values are shown above the branches, and bootstrap values (%) shown below the branches.

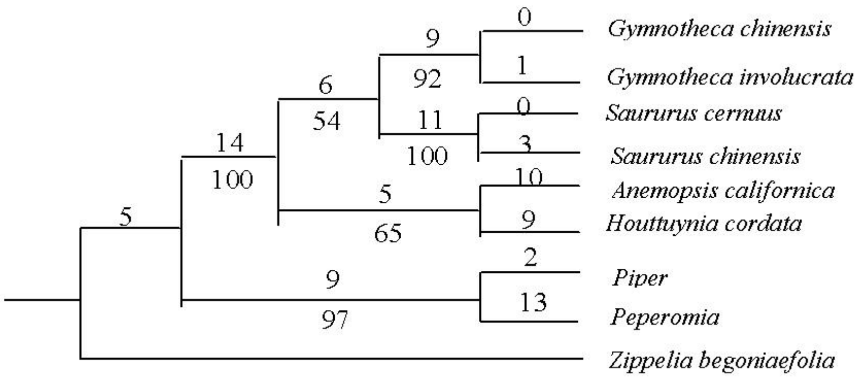

Fig. 5. The single parsimony tree of Saururaceae based on morphological data. Length = 97, CI =0.8351, RI = 0.7975, RC = 0.6659. Number of bootstrap replicates = 1000. Branch length values are shown above the branches, and bootstrap values (%) are shown below the branches.

the homoplasious characters. To reduce this interference, we should give more weight to the homologous characters in order to emphasize their effect. After weighting the characters according to re-scaled consistency indices (base weight = 2), we re-analyze the matrix once more with the same setting as before. “Base weight” is a degree of weighting. When analyzing a data matrix using equal weight, the base weight in all sites is equal to 1. For a phylogenetic program, it is the default status. In our study, we weight according to RC and set the base weight to be 2. It means that we weight homologous characters at 2 and the remaining homoplasious characters at 1.

A stable topology identical to Figure 5 is obtained. Length of the tree is 135 steps, CI = 0.9556, RI = 0.9439, RC = 0.902. Again, Saururaceae (100%), *Saururus* (100%), and *Gymnotheca* (96%) were monophyly. *Saururus* was the sister group of *Gymnotheca* (91%), and *Anemopsis* was the sister group of *Houttuynia* (92%).

5. Discussion

5.1. *Phylogeny of Saururaceae*

Let us first summarize the results from Section 4.

- We see from Figure 4 that the combined molecular data also strongly supports (1) the monophyly of Saururaceae (100%), *Saururus* (100%), and *Gymnotheca* (100%); (2) the sister group relationship between *Anemopsis* and *Houttuynia* (100%); (3) the sister group relationship between *Gymnotheca* and *Saururus* (100%); and (4) *A. californica* and *H. cordata* forming the first clade of Saururaceae.
- The trees inferred from separate DNA sequences of 18S (Figure 1), *trn*L-F (Figure 2), and *mat*R (Figure 3) also show the identical topology of Saururaceae, which is the same as the topology of Saururaceae from the combined DNA sequences.
- The molecular phylogenies also get strong support from the morphological analysis shown in Figure 5, which yields a topology identical to the tree from the combined DNA sequences.
- However, in the morphologically phylogenetic tree, the sister relationships between *Anemopsis* and *Houttuynia* (53%), and between *Gymnotheca* and *Saururus* (65%) are weak. Importantly, after weighting the characters according to RC indices, the sister relationships between *Anemopsis* and *Houttuynia* (92%), and between *Gymnotheca* and *Saururus* (91%) become strong.
- This result is surprising and differs from all the other phylogenetic opinions[484, 505, 631, 846, 908, 909] on Saururaceae.

Our results disagree with the systematic opinion of Wu and Wang[908, 909] on Saururaceae. However, our results partly agree with Wu,[906] who proposes that *Anemopsis* and *Houttuynia* are vicariant genera, and *S. chinensis* and *S. cernuus* are vicariant species. In a phylogenetic sense, vicariant genera or species may be interpreted as sister group. Our study well support the sister group relationships between *Anemopsis* and *Houttuynia*, and between *S. chinensis* and *S. cernuus*.

Our results are also not in consensus with Okada[631] and Lei *et al.*.[484] They suggest *Saururus* as the basal genus and *Anemopsis* and *Houttuynia* as “advanced” genera. Lei *et al.*[484] further suggest that *Gymnotheca* is the “most advanced.” Okada and Lei *et al.* separately construct a phylogeny of Saururaceae only according to the basic chromosome number of Saururaceae. However, according to Figure 6, the basic chromosome numbers in Saururaceae (character 57) is homoplasious and is not a dominant characterisitic for reconstructing the phylogeny of Saururaceae. Moreover, it appears difficult to conclude that extant groups such as *Circaeocarpus*, *Anemopsis*, *Gymnotheca*, *Houttuynia* are derived from another extant group (*Saururus*).

In terms of the sister relationship of *Anemopsis* and *Houttuynia*, our results partly agree with Tucker, *et al.*[846] who has generated a tree identical to the combined DNA sequence tree in our study; see Figure 4 of this chapter and Figure 5 of Tucker *et al.*.[846] Nevertheless, they treat *Saururus* as the first-derived genus in Saururaceae and believe that *Saururus* bear many plesiomorphies. The accepted tree in Tucker *et al.*[846] is supported with low bootstrap values.

We would like to address two points in Tucker *et al.*[846] One is the criterion for selecting out-group. *Cabomba*, *Magnolia*, *Chloranthus*, and even *Lactoris* and *Saruma* are not good out-groups for studying Saururaceae because they are too alien from Saururaceae according to the present understanding of angiosperm phylogeny.[27, 688, 786] Piperaceae is the best out-group of Saururaceae. The other is the interpretation of character 20 in Tucker *et al.*[846] on “whether a pair of stamens originated from separate primordia (0) or a common primordium (1)”. The stamens of *Houttuynia* are from separate primordia (0).[505, 841] However, this character is coded as 0 or 1 in Tucker *et al.*. When we correct it and re-analyze the same matrix using *Zippelia*, *Piper*, and *Peperomia* as out-groups, the topology is the same as Figure 5 in Section 4, and the bootstrap supports are high.

Liang[505] overweights a morphological character, “rhizomatous or not”, and hence supports the monophyly of Saururaceae. She treats “stoloniferous” and “separate initiation of bract-flower” as synapomorphies, and hence supports the sister relationship of *Gymnotheca* and *Anemopsis*. She also treats “common primordium initiation of bract-flower” as the synapomorphies of *Saururus* and *Houttuynia*. According to Figure 6, “stoloniferous or erect stem” (character 0) and “the ontogeny of bract-flower” in Saururaceae (character 33) are homoplasious and are not suitable to reconstruct phylogeny of Saururaceae as dominant characters. Therefore, our study does not support an overweight on a few particular morphological or ontogenetic characters.

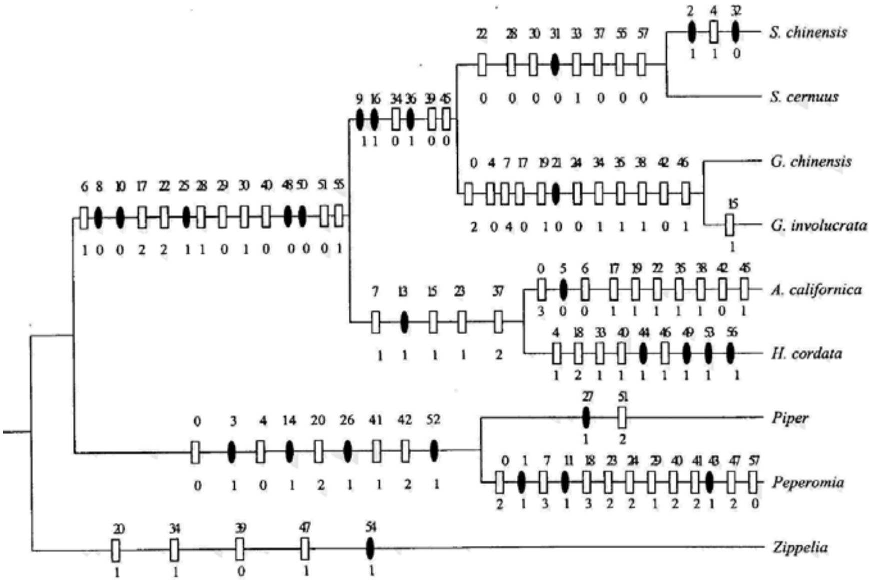

Fig. 6. The distribution of morphological characters. Character numbers are shown above the branches, and character states are shown below the branches. Black ovals represent homologous characters. Blank rectangles represent homoplasious or reversal characters. Homology is a similarity due to common evolutionary origin. Homoplasy is the existence of characters that have been subject to reversals, convergences, or parallelisms. Reversal is an evolutionary change whereby a character changes state and then changes back again. Convergence is an evolutionary event where the similarity between two or more characters is not inherited from a common ancestor. Parallelism is an evolutionary event where two identical changes occur independently. Convergences differ from parallelisms. In convergences, the ancestral characters are not the same. However, in parallelism, the ancestral characters are the same.

5.2. *The Differences Among Topologies from 18S, trnL-F, and matR*

The topologies from *trn*L-F (Figure 2) and *mat*R (Figure 3) are identical. In terms of arrangement of out-groups, the topology from 18S (Figure 1) slightly differs from the topology from *trn*L-F and *mat*R. Two points may cause the different arrangements of the out-groups. Firstly, we use different out-groups in separate analysis of different gene sequences. Besides *Z. begoniaefolia*, the out-groups are *Pe. serpens* and *P. mullesua* in the analysis of 18S sequences, *Pe. tetraphylla* and *P. mullesua* in the analysis of *trn*L-F sequences, and *Peperomia obtusifolia* and *Piper betle* in the analysis of *mat*R sequences. Secondly, Piperaceae is a large family, and so the arrangement of out-groups should be identical if we use more species of *Piper* and *Peperomia*.

Nevertheless, for the study on Saururaceae, the treatment for out-group in this chapter is suitable. Here, the main out-groups are *Z. begoniaefolia*, *Pe. tetraphylla* and *P. mullesua*. For the fast-mutating *trn*L-F, we use these three out-groups. For the slow-mutating functional 18S and *mat*R genes, we replace *Pe. tetraphylla* and *P. mullesua* with close species because the sequences of these close species are available in GenBank. Actually, the 18S sequences between *Pe. tetraphylla* and *Pe. serpens* are almost identical. So, in the phylogenetic analysis of 18S sequences of Saururaceae, it is suitable to replace *Pe. tetraphylla* with *Pe. serpens*. Similarly, in the analysis of *mat*R sequences, it is suitable to replace *P. mullesua* with *P. betle* and to replace *Pe. tetraphylla* with *Pe. obtusifolia*. Of course, support values should be higher if the identical out-groups are used.

The support values for each branch of Saururaceae based on 18S, *trn*L-F, and *mat*R are slightly different, and a few of them are quite low (Figures 1–3). The reasons are as follow. Firstly, since different genes have different characters, the phylogenetic trees based on different genes are different. Secondly, since sequences are from different sources, and thus have different systematic errors, the phylogenetic trees based on different sequences are different. In particular, the 18S sequence of *A. californica* and the *mat*R sequences of *A. californica*, *H. cordata*, and *S. cernuus* are from GenBank; but the other sequences are from the Laboratory for Plant Biodiversity and Biogeography, Kunming Institute of Botany, Chinese Academy of Sciences.

In conclusion, different out-groups, different genes, and different laboratory systems result in the slight difference among separate phylogenetic topologies. However, the difference is so small that it does not affect the topologies greatly. Moreover, the combination of DNA sequences reduces the differences among separate DNA sequence alignments (Figure 4).

5.3. *Analysis of Important Morphological Characters*

Some authors[484, 505, 631, 846, 908, 909] think that the ancestral Saururaceae is similar to the extant *Saururus*, which has free carpels, free stamens and superior ovaries. Moreover, they believe that six stamens (character 18), free stamens (character 19), hypogynous stamens (character 17), four carpels (character 23), superior carpels (character 22), free carpels (character 24), freedom of stamens and carpels (character 21), and marginal placenta (character 28) are primitive features of *Saururus*. However, these primitive characters can be interpreted as synplesiomorphies in most cases and are not important for phylogenetic reconstruction.[349] As shown in Figure 6, characters 18, 19, 23, 24 are homoplasious, and characters 17, 22, 28 are reverse in Saururaceae. These characters should not be used as dominant factors when reconstructing phylogeny of Saururaceae. In all analysis of this study, Figures 1–6, *Saururus* appears not to be the first derived genus in Saururaceae.

Free-carpel has long been regarded as a relictual feature in angiosperms.[79,379] However, in the case of Saururaceae, according to the out-group comparison with *Zippelia*, *Piper*, and *Peperomia* in Figure 6, free-carpel (character 24) should be recognized as a homoplasious character. So should more-stamens (character 18), free-stamens (character 19), and more-carpels (character 23) be thought as homoplasious characters. Hypogynous-stamen (character 17), superior-ovary (character 22), and marginal-placenta (character 28) are regarded as primitive characters in phylogenetic reconstruction. However, they are reverse in Saururaceae. All of the above characters should not be dominant characters when reconstructing phylogeny of Saururaceae. Similar situation occurred in *Archidendron* (*Leguminosae*). Taubert[821] and following authors put *Archidendron* on a primitive position because *Archidendron* has several ovaries. But evidences from flowers, pollens, and wood anatomy, as well as the whole sequence of specialization from Ceaesalpinieae to Ingeae, indicate that *Archidendron* is highly advanced and its immediate ancestor has single ovary.[671]

6. Suggestions

Finally, let us note in the following subsections some of the skills that are important when reconstructing phylogeny.

6.1. *Sampling*

If a different set of sample is used in the study, we can expect the results to be different. For example, the tree produced from the 18S DNA sequences is slightly different from the tree produced from morphological data, mainly on the positions of out-groups and bootstrap supports. Concretely, *Piper* diverges earlier than *Peperomia* in the tree from the 18S DNA sequences in Figure 1, but *Peperomia* is the sister group of *Piper* in the tree from morphological data in Figure 5.

What is the reason? In the analysis of the 18S DNA sequences, the out-group *Piper* includes only *P. mullesua*; and *Peperomia* includes only *Pe. Serpens*. However, in the analysis of morphological data, the out-group *Piper* includes all species of the whole genus, and so does *Peperomia*. This difference in sampling causes the slight difference in the resulting topologies. So it is better to sample all species if the studied taxon is small, and to sample as many and as representative as possible if the studied taxon is big.

6.2. *Selecting Out-Group*

It is common that different out-group gives different result. What is the best out-group for a studied taxon? Hennig[349] points out that the sister group of a taxon is the best out-group, and one of the main tasks of phylogenetic analysis is to look for the sister group. In our case, identifying the sister group of Saururaceae becomes the critical procedure of the whole analysis.

In order to look for the sister group of Saururaceae, we have checked many classical works. Hutchinson[379] and Cronquist[179] both put Piperaceae, Saururaceae, and Chloranthaceae in Piperales. Melchior—see pages 764–796 of Brummitt's famous volume[113]—circumscribes Saururaceae, Piperaceae, Chloranthaceae, and Lactoridaceae in Piperales. In the systems of Dahlgren,Thorne, and Takhtajan—see respectively pages 777–783, 770–776, and 790–799 of Brummitt's famous volume[113]—Piperales only includes Saururaceae and Piperaceae although Takhtajan[815] separates Peperomiaceae from Piperaceae. Chase[139] backs the sister relationship between Piperaceae and Saururaceae in an analysis for *rbc*L sequences. So does Qiu *et al.*[688] for *rbc*L, *atpB*, 18S, *mat*R and *atp1* from three genomes, and Hoot[363] and Soltis *et al.*[786] for *atpB*, *rbc*L and 18S.

In conclusion, according to not only classical morphology systematics but also molecular systematics,[27, 688, 786, 846, 907] it is clear that Piperaceae is the sister group of Saururaceae. Thus Piperaceae is the best out-group when studying Saururaceae.

6.3. *Gaining Sequences*

Mistakes in sequences are not corrected even by the best analysis methods. So it is important to ensure that the sequences are accurate. How to guarantee accurate sequences? Firstly, the experimental skills and the experimental system should be good. Secondly, check each sequence with its electrophoretic map using SeqEd program (Applied Biosystems) or other softwares. Thirdly, determine the ends of sequences by comparing with similar sequences. *E.g.*, in our case, the ends of 18S gene are determined by comparing with the sequences of AF206929. It is more convenient to directly cite sequences from GenBank,[77] EMBL, or other databases. But some sequences and other data in these databases are rough. So better check the downloaded data before using them.

6.4. *Aligning*

When aligning sequences using Clustal-X, MEGA2b3, or other softwares, different input ordering of sequences can result in different alignment matrices, and consequently different phylogenetic trees. It is better to input high-quality and similar sequences first when using these softwares. When DNA are introns or gene spacers, alignment parameters should be small. When DNA are functional genes, alignment parameters should be big. Generally, the values should be bigger when genes are more conservative.

6.5. *Analyzing*

When reconstructing phylogeny, distance matrix methods, maximum parsimony methods, maximum likelihood methods, and methods of invariants are frequently used. No method can resolve all problems in all cases. So it is necessary to know which method should be used in different cases.

Different methods make different assumptions. Maximum parsimony methods assume that the evolutionary process is parsimonious. Hence it considers a tree with fewer substitutions or homoplasy as better than a tree with more substitutions or homoplasy. When the divergence between characters or sequences is small, maximum parsimony methods work well. In contrast, when the divergence is large, they work badly. Particularly, the parsimony methods are easily mislead when analyzing sequences that are evolving fast.

The unweighted pair-group with arithmetic mean method (UPGMA) assumes a constant rate in all branches. This assumption is often violated and thus the UPGMA tree often has errors in branching order.

Other distance methods assume that unequal rates among branches can be corrected by distances in a distance matrix. The performance of correction is affected by accuracy of the distances estimated. Normally, when the distances are small or the sequences of DNA or protein are long, the distance methods work well. However, when the sequences are short, distances are large or the differences of rates among sites are large, distance methods work badly.

Maximum likelihood methods make clear assumption on evolutionary rate or the base substitution model. Using maximum likelihood methods, we can develop a program with options for different evolutionary rates or different base substitution models. However, maximum likelihood methods need too much computation. So they are difficult to run fast in today's computer systems when the operational taxonomic units (OTU) are many or the sequences are long.[501]

The computational time is also different in different methods. In general, distance methods take the least time, and maximum likelihood methods take the most time. When reconstructing phylogeny, the common search techniques used are heuristic, branch-and-bound, and exhaustive. Heuristic search is rough but fast. Exhaustive search is accurate but slow, and its computational need is tremendous.

Branch-and-bound search is in the middle. Generally, use exhaustive search when the number of OTU is small, especially less than 11. Use heuristic search when the number of OTU is very large. Use branch-and-bound search when the number of OTU is not very large.

We can use many different programs to reconstruct phylogeny, such as PAUP, PHYLIP, MEGA, MacClade, WINCLADA, PAML and DAMBE. Although different programs have their own virtues, the most popular are PAUP and PHYLIP. Browse `http://evolution.genetics.washington.edu` to view some of these phylogenetic softwares.

For molecular phylogenetics, it is better to reconstruct phylogeny using as many different types of DNA sequences as possible. Analysis of separate DNA sequence matrices should be under the same setting. The separate DNA sequence matrices can also be combined into a larger matrix, and then analyze under the same setting. Generally, a combined matrix is better than separate matrices; because the combined matrix has more informative sites, and thus is more helpful to reconstruct a reliable phylogeny.

6.6. *Dealing with Morphological Data*

There are many articles on how to reconstruct phylogeny using morphological data,[255] so we only address two points here. Firstly, morphological characters should be polar and unordered. Typically, 0 is taken to represent the most primitive when numbering a character series, and the other positive integers are taken to represent different numbers and do not represent the evolutionary order. It means that, for example, the evolutionary step from 0 to 1 is one, and the evolutionary step from 0 to 4 is also one. Secondly, the values numbering out-groups' characters have not only 0, but also the other positive integers, such as 1, 2, 3, and 4.

6.7. *Comparing Phylogenies Separately from Molecular Data and Morphological Data*

A trend in phylogenetics is to combine matrices from DNA sequences and morphological data in order to make a more comprehensive analysis. [516] It is a reasonable method if a trade-off can be obtained between the two types of data. However, because a DNA sequence matrix is much longer and have much more informative sites than a morphological matrix, the information of the DNA sequences often overwhelms the information of the morphological data when analyzing a combination of the two types of data. Of course, we can enlarge the effect of the morphological matrix by weighting it. However, what is the criterion for the extra weight? How much extra should the weight be?

In our opinion, It is easier to compare the phylogenetic tree from molecular data with the phylogeny from morphological data. It is ideal if all of the phylogenies based on different data from variable subjects are identical. However, differences and discords among results from different data often occur. In our case, the phylogenies of Saururaceae from molecular data and morphological data are identical. However, they are almost opposite to the traditional opinions on the phylogeny of Saururaceae. Traditional opinions assert that *Saururus* is the most primitive genus in Saururaceae. However, we regard the *Saururus-Gymnotheca* clade as the sister group of the *Anemopsis-Houttuynia* in Saururaceae. What is the matter? We have to check. It costs much time because we have to check every original datum and every procedure of analysis. At last, we find out the reasons that cause the difference between our result and others' results.

6.8. *Doing Experiments*

Even though the conclusions of our case study are reasonable, they are based only on deduction. So, it is necessary to confirm our results by doing wet experiments. In order to check whether the concluded homoplasies of some morphological characters are true, we should design and perform experiments on floral development and histochemistry of Saururaceae and Piperaceae. Similarly, if it is possible, one should confirm one's conclusions by wet experiments after reconstructing phylogeny.

Acknowledgements

We would like to express our hearty thanks to De-Zhu Li, Han-Xing Liang, Shirley C. Tucker, Andrew W. Douglas, and Louxin Zhang for helpful suggestions; to Jun-Bo Yang, Zhen-Hua Guo, Yong-Yan Chen, Feng Wang, and Haiquan Li for their valuable assistance.

	characters	character states
0	Growth habit	wood or liana(0), erect herb(1), stolon(2), lotiform plant(3).
1	Leaves position	alternate(0); alternate, lumbricine or whorled(1).
2	Terminal leaf of stem in reproductive period	green(0), white(1).
3	Stipule	adnation with a stipe of a leaf(0), not present(1).
4	Tomentum on lamina	no tomentum(0), tomentum on underside(1), tomentum on both sides(2).
5	Leaf venation	pinnate venation(0), hyphodromous(1), palmate venation(2).
6	Lateral leaf venation	no lateral leaf venation(0), dichotomous(1), not dichotomous(2).
7	Areoles	areolation lacking(0), incomplete(1), incomplete or imperfect(2), imperfect(3), imperfect or perfect(4).
8	Number of stem vascular cylinder	1(0), 2(1).
9	Fibre in stem	absent(0), discontinuous(1), continuous(2).
10	Perforation plate type in vessel members	scalariform perforation plate(0), simple perforation plate(1).
11	Number of inflorescence at one site	one(0), many(1).
12	Floral symmetry	radial symmetry(0), dorsiventral or zygomorphic symmetry(1).
13	Peloria	no abnormal regular flower(0), abnormal regular flower present(1).
14	Shape of floral bracts	lanceolate(0), peltate(1).
15	Color of inflorescence involucrum	green(0), showy(1).
16	Flower-bract stalk	no stalk(0), flower-bract stalk present(1).
17	Stamens position	hypogynous(0), perigynous(1), epigynous(2).
18	Number of stamens	6(0), 4(1), 3(2), 2(3).
19	Stamen fusion	free(0), connate(1).
20	Anther dehiscence	stomium along entire length of anther(0), stomium predominantly in proximal position(1), in distal position(2).

Fig. 7. Morphological characters and their states.

	characters	character states
21	Adnation of stamens and carpels	free(0), fused partly(1).
22	Ovary position	superior ovary(0), perigynous ovary(1), inferior ovary(2).
23	Number of carpels	4(0), 3(1), 1(2).
24	Carpels adnation	free(0), fused(1), single carpel(2).
25	Style presence	no style(0), style present(1).
26	Stigma shape	stigmatic stylar cleft(0), capitate or tufted(1), divided stigma(2).
27	Ovule number per carpel	1(0), $<$ 1(1).
28	Placenta	marginal placenta(0), parietal placenta(1), basal placenta(2).
29	Ovules per carpel	$\geq$ 3(0), 1(1), $<$ 1(2).
30	Number of carpel vascular bundle	2(0), coadnate(1), 1(2).
31	Vascular bundle fusion of stamens and carpels	free(0), fused partly(1).
32	Fusion of Adaxial and abaxial carpel bundle	free(0), fused partly(1).
33	Genesis of bract-flower	discrete bract and flower intition(0), common primordial inition(1).
34	Genesis order of carpels	middle primordium first(0), bilateral primordium first(1), appear simultaneous or single or common primordium(2).
35	Genesis of stamens	discrete primordium(0), common primordium(1).
36	Genesis order of stamens	bilateral stamens first(0), middle stamens first(1).
37	Genesis pattern of median sagittal stamens	in pair(0), adaxial axis first(l), no adaxial or abaxial stamen(2).
38	Genesis pattern of bilateral stamen pair	discrete primordium(0), common primordium(1).
39	Median sagittal carpels	adaxial and abaxial carpels(0), adaxial(1).
40	Germinal aperture	anasulcate(0), anasulcate and anatrichotomosulcate(1), inaperturate(2).

Fig. 8. Morphological characters and their states (continued).

	characters	character states
41	Ornamentation of pollen exine	foveolae(0), verruculose(1), large verruculose(2).
42	Small verruculose at the edge of foveolae of pollen tectate	absent(0), present(l), narrow belt of granule on tectate(2).
43	Number of microsporosac	4(0), 2(1).
44	Microspore gennesis	simultaneous(0), successive(1).
45	Type of minor tetrad	bilateral symmetry, T-shape and cross-shape(0), bilateral symmetry and cross-shape(1), bilateral symmetry(2).
46	Pollen abortion	no abortion(0), abortion(1).
47	Layers of integument	two layers(0), outer layer degradation(1), only inner layer(2).
48	Micropyle	consist of inner and outer integument(0), consist of inner integument(1).
49	Nucellus	crassinucellate ovule(0), tenuinucellate ovule(1).
50	Functional megaspore	from one of megaspore tetrad(0), from four of megaspore tetrad(1).
51	Type of embryo sac	polygonum type(0), drusa or peperomia type(1), fritillaria type(2).
52	Fusion of two central nuclei of embryo sac	before fertilization(0), after fertilization(1).
53	Apomixis	apomixis absent(0), apomixis present(1).
54	Perisperm type	cellular type perisperm(0), nuclear type perisperm(1).
55	Fruit type	folicule(0), capsule(1), berry(2).
56	Ploid	biploid(0), polyploid(1).
57	Basic number of chromosome	11(0), other number(1).

Fig. 9. Morphological characters and their states (continued).

```
                            111111111122222222223333333333444444444455555555
Taxon/Node        0123456789012345678901234567890123456789012345678901234567

S. chinensis      1010121201001000100000000100000000100100001000000000000000
S. cernuus        1000221201001000100000000100000000110010000010000000000000
G. chinensis      2000021401001000120101201100101110111110000000100000000101
G. involucrata    2000021401001001120101201100101110111110000000100000000101
A. californica    3000200?0?000101010101111100101110210211000001000000000101
                              1
H. cordata        1000121102000101022001211100101111200201101012100100010111
                              1
Z. begoniaefolia  10002222121010000000110010002221101001001010020110 11?01201
                                                              1
Piper             000100001210101000002000101122221?020010111?002001012100201
                      1221          1  1 1  2                           1 1
                         2
Peperomia         210101231011101000302102201021 21?0200?0?222102021011?00200
                      22   2                                              1
```

Fig. 10. The matrix of the morphological characters.

CHAPTER 12

FUNCTIONAL ANNOTATION AND PROTEIN FAMILIES: FROM THEORY TO PRACTICE

Noam Kaplan

The Hebrew University of Jerusalem
kaplann@pob.huji.ac.il

Ori Sasson

The Hebrew University of Jerusalem
ori@cs.huji.ac.il

Michal Linial

The Hebrew University of Jerusalem
michall@cc.huji.ac.il

We discuss two bioinformatics tools, ProtoNet and PANDORA, that deal with different aspects of protein annotations and functional predictions. ProtoNet uses an approach of protein sequence hierarchical clustering to detect remote protein relatives. PANDORA uses a graph-based method to interpret complex protein groups through their annotations.

ORGANIZATION.

Section 1. We introduce the objectives and challenges of computationally inferring function from sequence information. We discuss the shortcomings of some commonly used tools for this purpose. ProtoNet and PANDORA are two tools designed to complement these commonly used tools and alleviate their shortcomings.

Section 2. Then we present ProtoNet. ProtoNet is centered around the concept of homology transitivity. ProtoNet clusters protein sequences into a hierarchy based on sequence similarity and homology transitivity. Thus the hierarchy parallels the evolutionary history of these protein sequences. We then illustrate—using Histone proteins—the application of ProtoNet to detect remote protein relatives.

Section 3. We describe ProTarget, which is built on top of ProtoNet. ProTarget is a useful tool for selection of protein targets that have a high probability of exhibiting to a new fold.