Lecture Notes in Mathematics

Avner Friedman (Ed.)

Tutorials in Mathematical Biosciences IV

Evolution and Ecology

1922

Mathematical Biosciences Subseries

Springer

mbi
tutorials

Avner Friedman (Ed.)

Tutorials in Mathematical Biosciences IV

Evolution and Ecology

With Contributions by:

C. Cosner · D. Janies · L.S. Kubatko
Y. Lou · T. Nagylaki

Springer

mbi
Mathematical Biosciences Institute
at The Ohio State University

Editor

Avner Friedman
Mathematical Biosciences Institute
Ohio State University
231 West 18th Avenue
Columbus, OH 43210-1292
USA

e-mail: afriedman@math.ohio-state.edu
afriedman@mbi.ohio-state.edu

Library of Congress Control Number: 2007933684

Mathematics Subject Classification (2000): 35J20, 35J60, 35K55, 35K57, 62P10, 62P12, 92B99, 92D10, 92D15, 92D25, 92D40

ISSN print edition: 0075-8434
ISSN electronic edition: 1617-9692
ISBN 978-3-540-74328-6 Springer Berlin Heidelberg New York
DOI 10.1007/978-3-540-74331-6

This work is subject to copyright. All rights are reserved, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilm or in any other way, and storage in data banks. Duplication of this publication or parts thereof is permitted only under the provisions of the German Copyright Law of September 9, 1965, in its current version, and permission for use must always be obtained from Springer. Violations are liable for prosecution under the German Copyright Law.

Springer is a part of Springer Science+Business Media
springer.com
© Springer-Verlag Berlin Heidelberg 2008

The use of general descriptive names, registered names, trademarks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

Typesetting by the author and SPi using a Springer LaTeX macro package

Cover design: *design & production* GmbH, Heidelberg

Printed on acid-free paper SPIN: 12109630 41/SPi 5 4 3 2 1 0

Preface

This is the fourth volume in the series "Tutorials in Mathematical Biosciences." These lectures are based on material which was presented in tutorials or developed by visitors and postdoctoral fellows of the Mathematical Biosciences Institute (MBI), at The Ohio State University. The aim of this series is to introduce graduate students and researchers with just a little background in either mathematics or biology to mathematical modeling of biological processes. The first volume was devoted to mathematical neuroscience, which was the focus of the MBI program 2002–2003. The second volume dealt with mathematical modeling of calcium dynamics in signal transduction, the focus of the MBI program in the winter of 2004. The third volume dealt with topics of cell cycle, tumor growth, and cancer therapy; these topics featured in several workshops held at the MBI in the fall of 2003. The present volume deals with a variety of topics of evolution and ecology, which were considered in the MBI during the year 2005–2006. These topics include phylogenetics; evolution of genes through migration–selection; ecological modeling; and evolution of dispersal and population dynamics. Documentation of the 2005–2006 activities, including streaming videos of the workshops, can be found on the Web site: http://mbi.osu.edu.

Phylogenetics is the study of the evolutionary relations of genes and organisms. Phylogenetic trees are represented by graphs in which the leaves represent observed biological entities. In constructing such graphs, one tries to trace the evolution of species, traits, or diseases. The first two chapters of this volume deal with phylogenetics. Chapter 1 is a general survey on estimation of phylogenetic trees with emphasis on likelihood methods. Chapter 2 is concerned with computational methods of very large trees, exploring other optimality methods, with application to the study of the evolution of SARS and influenza.

The next three chapters deal with population genetics and population dynamics. Chapter 3 introduces reaction–diffusion equations as a mathematical framework to study ecological models. It then addresses the following ecological questions: what is the minimal patch size necessary to support a population?; when do biological invasions occur?; and what spatial patterns can form?

Chapter 4 focuses on evolution and genes. The genetic composition of a population is described by genotypic or allelic frequencies, using either deterministic models or stochastic models. The models presented here are both discrete and continuous. The questions discussed include the loss, or the maintenance, of a specified allele, and the stability of completely polymorphic equilibria.

The final chapter is concerned with the effects of dispersal and spatial heterogeneity on population dynamics, via reaction–advection–diffusion models. Issues regarding how advection along resource gradients affect the extinction of species or how invasion of rare species may take place are considered.

It is not uncommon to see the same biological processes benefit by using different mathematical and statistical approaches. This volume is a good example: Although the mathematical and statistical tools developed or reviewed here are quite varied, the biological themes have a common thread as they all deal with the evolution of species in an evolving ecological system.

I express my appreciation and thanks to Daniel Janies, Diego Pol, Laura Salter-Kubatko, Thomas Nagylaki, Yuan Lou, and Chris Cosner for their marvelous contributions. I hope this volume will serve as a useful introduction to those who want to learn about important and exciting problems that arise in evolution and ecology.

此为试读，需要完整PDF请访问：www.ertongbook.com

Contents

1

Inference of Phylogenetic Trees

L.S. Kubatko

Departments of Statistics and Evolution, Ecology, and Organismal Biology,
The Ohio State University, Columbus, OH 43210, USA
email: lkubatko@stat.osu.edu

Study of the evolutionary relationships among organisms has been of interest to scientists for over 100 years. The earliest attempts at inferring evolutionary relatedness relied solely on observable species characteristics. Modern molecular techniques, however, have made available an abundance of DNA sequence data, which can be used to study these relationships. Today, it is common to consider the information contained in both types of data in order to obtain robust estimates of evolutionary histories.

These evolutionary histories are most commonly represented by a phylogenetic tree, which is mathematically described as an acyclic connected graph (V, E), where V is the set of vertices and E is the set of edges. Vertices connected through only a single edge are called terminal nodes, while vertices connected by more than one edge are called internal nodes. In phylogenetic tree reconstruction, it is common to assume that trees are bifurcating, so that each internal node is connected through exactly three edges, with the exception that for a rooted tree the root is connected through two edges.

Estimation of the phylogenetic relationships among a collection of organisms given genetic data for these organisms can be divided into two distinct problems. The first is to define the particular criterion by which we compare the fit of a particular phylogenetic hypothesis to the observed data. The second is to search the space of possible phylogenies for the particular tree or trees that provide the best fit to the data. In this chapter, we give an overview of these two problems, with particular emphasis on the maximum parsimony and maximum likelihood criteria for comparing trees. Techniques for searching the space of trees for optimal phylogenies under these criteria are also discussed. Throughout the chapter, we use two data sets to illustrate the main ideas. We begin by defining some of the commonly used terminology, and by providing a careful description of the data used in phylogenetic analysis.

1.1 Introduction and Terminology

1.1.1 Phylogenetic Trees

As described earlier, a phylogenetic tree can be viewed as a graph for which the terminal nodes represent organisms for which data are observed, called *taxonomic units* or *taxa*, while the internal nodes represent hypothetical ancestral organisms. The edges connecting the nodes are generally referred to as branches and denote ancestry-descent relationships. Often, the lengths of the branches are taken to represent evolutionary time. In this chapter, the word *topology* will be used to refer to the labeled branching pattern of a tree without regard to branch lengths.

Phylogenetic trees are called rooted when the location of the common ancestor of all the taxa in the tree is identified, or unrooted when no such common ancestor is specified. Rooted phylogenetic trees may or may not satisfy the assumption of a molecular clock. The molecular clock hypothesis is that the rate of evolution is approximately constant over time. When all of the sequences in the tree are contemporaneous, this assumption restricts the lengths of the branches so that the sum of the branch lengths connecting each taxon to the root is the same for all taxa. Examples of phylogenetic trees are shown in Fig. 1.1.

As the number of taxa under consideration grows, the number of distinct topologies increases rapidly. For n taxa, the number of unrooted labeled bifurcating topologies is

$$\prod_{i=3}^{n}(2i-5). \tag{1.1}$$

An unrooted topology has $n-2$ internal nodes and $2n-3$ branches. Because adding a root to an n-taxon tree amounts to placing it along any of the $2n-3$ branches, the number of rooted topologies for n taxa is the found by applying (1.1) for $n+1$. A rooted topology for n taxa contains $n-1$ internal nodes and $2n-2$ branches. Table 1.1 shows the rapid increase in the number of topologies as a function of the number of taxa.

Table 1.1. Number of topologies, internal nodes, and branch lengths for unrooted bifurcating topologies as a function of the number of tips in the tree

Number of tips	Number of topologies	Number of internal nodes	Number of branches
5	15	3	7
10	2,027,025	8	17
20	2.2164×10^{20}	18	37
50	2.8381×10^{74}	48	97

Table 1.2. DNA sequences for a portion of the *L1* gene for seven Group A9 human papillomaviruses

HPV16	ATGTGGCTGCCTAGTGAGGCCACTGTCTACTTGCCTCCTGTCCAGTATCTAAGGTTG
HPV35h	ATGTGGCGGTCTAACGAAGCCACTGTCTACCTGCCTCCAGTTCAGTGTCTAAGGTTG
HPV31	ATGTGGCGGCCTAGCGAGGCTACTGTCTACTTACCACCTGTCCAGTGTCTAAAGTTG
HPV52	ATGTGGCGGCCTAGTGAGGCCACTGTGTACCTGCCTCCTGTCCTGTCTCTAAGGTTG
HPV33	ATGTGGCGGCCTAGTGAGGCCACAGTGTACCTGCCTCCTGTCCTGTATCTAAAGTTG
HPV58	ATGTGGCGGCCTAGTGAGGCCACTGTGTACCTGCCTCCTGTCCTGTGTCTAAGGTTG
RhPV1	ATGTGGCGGCCTAGTGACTCCAAGGTCTACCTACCACCTGTCCTGTGTCTAAGGTGG

See Sect. 1.1.3 for a description of the data.

1.1.2 Data for Phylogenetic Estimation

The most common type of data used in phylogenetic inference is discrete character data. These data can be represented by a matrix X in which entry x_{ij} represents the particular state of the character observed for taxon i at position j. Both morphological data and molecular sequence data (DNA, RNA, amino acid, or protein sequences) are examples of discrete character data. For data of this type, we assume that each character (column) in the data matrix is homologous, which means that in each of the taxa the particular state observed was derived from an ancestral state that was common to all of the taxa in the data matrix. In practice, assessment of homology can be difficult, particularly for molecular sequence data.

Development of a data matrix for use in phylogenetic inference for molecular sequence data is a nontrivial task, because the molecular sequences are derived individually for each taxon and must subsequently be placed into a data matrix so that the assumption of homology is likely to be satisfied. The process of constructing the data matrix for a collection of taxa is called *sequence alignment*. Table 1.2 shows an example of an aligned portion of the *L1* gene for seven human papillomaviruses. In this chapter, the problem of sequence alignment will not be discussed, and we will assume that the data have already been aligned. The interested reader is referred to several references on the topic: [54, 64, 71, 82, 83].

1.1.3 Example Data Sets

Throughout this chapter, we will use two data sets to illustrate several techniques for phylogenetic inference. The first is a set of viral sequences for a particular gene, and the second consists of both morphological and molecular data on cephalopods. Further details concerning each data set are given below.

Papillomaviruses

Papillomaviruses are a group of viruses that infect a variety of organisms ranging from birds to mammals, including humans. They are small nonenveloped

DNA viruses that generally cause benign epithelial lesions, though some types may cause malignancies. The papillomavirus genome is approximately 8,000 base pairs in length and is divided into an early region (E), which encodes genes expressed immediately after infection of the host, and a late region (L), which encodes two capsid proteins. The early region comprises over 50% of the genome, and contains six open reading frames (E1, E2, E4, E5, E6, and E7). The late region comprises approximately 40% of the genome and encodes two proteins, L1 and L2. The remaining 10% of the genome is a long control region (LCR) that does not code for proteins but does contain transcription factor binding sites and the origin of replication.

Papillomaviruses are classified into types, subtypes, and variants based on the sequence of the *L1* gene. They are also grouped based on sequence similarity, host type, and pathogenic characteristics. In this example, we consider the sequence of the *L1* gene for thirty Group A papillomaviruses, 28 of which infect humans. Notable among this collection of sequences are human papillomavirus (HPV) types 16, 18, and 31, which are found to be associated with over 95% of cervical cancer cases [87]. The particular sequences studied here, as well as their genetic subtype and pathology, are listed in Table 1.3. More information on these particular sequences can be found in Ong et al. (1997), and information concerning the genetics of papillomaviruses in general can be found in Zheng and Baker (2006).

For this example, aligned DNA sequences were downloaded from the HPV Database maintained by Los Alamos National Labs (http://hpv-web.lanl.gov/). This alignment was edited by limiting the analysis to only the 30 taxa in Table 1.3, removing the sequence prior to the start codon in all taxa, and removing all sites for which all of the taxa had an insertion or deletion, resulting in 1,560 aligned sites.

Cephalopods

Cephalopods (e.g., squids, cuttlefishes, octopi) are a diverse class of molluscs containing over 800 species. They inhabit a wide range of marine environments, from coastal to benthic waters, and vary in size from 10 mm to several meters. Taxonomically, the class is divided into two groups, Nautiloidea and Coleoidea. Nautiloidea contains only a single genera, while Coleoidea contains all remaining extant taxa. Three subgroups within Coleoidea are recognized: Decabrachia (squids and cuttlefishes), Octobrachia (octopi), and Vampyromorpha. The placement of Vampyromorpha has been controversial, with some analyses supporting a sister relationship with Octobrachia and others placing Vampyromorpha with Decabrachia [6, 8].

For this example, we consider a subset of the data examined by Lindgren et al. [46], which includes both molecular and morphological data for 78 molluscs. Fifteen taxa, including representative taxa from the Decabrachia and

Table 1.3. Group A papillomaviruses, genetic subtypes, risk classification [38], and host tissue infected for the virus types studied here

Genetic subtype	Group	Risk classification	Host tissue infected
HPV32	A1	Low	Oral
HPV42	A1	Low	Genital
HPV3	A2	Low	Cutaneous
HPV10	A2	Low	Cutaneous
HPV2a	A4	Low	Cutaneous, mucousal
HPV27	A4	Low	Cutaneous, genital
HPV57	A4	Ambiguous	Oral, genital
HPV26	A5	Ambiguous	Cutaneous, possibly genital
HPV51	A5	High	Genital
HPV30	A6	Low	Cutaneous, mucousal
HPV53	A6	Ambiguous	Genital
HPV56	A6	High	Genital
HPV18	A7	High	Genital
HPV45	A7	High	Genital
HPV39	A7	High	Genital
HPV59	A7	High	Genital
HPV7	A8	Low	Cutaneous, oral
HPV40	A8	Low	Genital
HPV16	A9	High	Genital
HPV35h	A9	High	Genital
HPV31	A9	High	Genital
HPV52	A9	High	Genital
HPV33	A9	High	Genital
HPV58	A9	High	Genital
RhPV1	A9	Unclassified	Genital
HPV6b	A10	Low	Oral, genital
HPV11	A10	Low	Oral, genital
HPV13	A10	Low	Oral, genital
PCPV1	A10	Unclassified	Oral
HPV34	A11	Low	Oral, genital

Octobrachia as well as Vampyromorpha, are considered here, for both the morphological data assembled by Lindgren et al. (2004) and for three nuclear genes, 18S (3,477 sites, of which 808 are parsimony informative), 28S rRNA (667 sites, of which 238 are parsimony informative), and histone H3 (327 sites, of which 73 are parsimony informative), that they examined. The taxa selected for analysis are shown in Table 1.4. This data set will be used to highlight the differences in analyzing molecular and morphological data in a phylogenetic context.

此为试读，需要完整PDF请访问：www.ertongbook.com

Table 1.4. Cephalopod taxa included in our examples (from Lindgren et al. (2004))

Group	Species name	Type of cephalopod
Octobrachia		
	Stauroteuthis syrtensis	Cirrate octopus
	Thaumeledone guntheri	Benthic octopus
Vampyromorpha		
	Vampyroteuthis infernalis	Vampire quid
Decabrachia		
	Sepia officinalis	Cuttlefish
	Heteroteuthis hawaiiensis	Bobtail squid
	Spirula spirula	Ram's horn squid
	Idiospeius pygmaeus	Pygmy squid
	Loligo pealei	Common market squid
	Architeuthis dux	Giant squid
	Enoploteuthis leptura	Open ocean squid
	Pyroteuthis margaretifera	Open ocean squid
	Gonatus fabricii	Open ocean squid
	Histioteuthis hoylei	Open ocean squid
	Ommastrephes bartrami	Open ocean squid
	Psychroteuthis sp.	Open ocean squid

1.2 Optimality Criteria

Given a data matrix X consisting of either aligned molecular sequences or morphological data, it is necessary to develop methods for constructing a phylogenetic tree that appropriately represents the information concerning evolutionary relationships contained in X. There are three general classes of methods for constructing phylogenies from a given data matrix. The first set of methods are distance methods, in which the original data matrix X is first converted to a matrix of pairwise distances between taxa, and these distances are used to construct the phylogeny. Distance methods will not be considered further here, but see [45, 55, 68, 77] for details.

The second two methods, parsimony and maximum likelihood, are based on the definition of a criterion for comparing alternative trees. The problem of constructing a phylogeny from a data matrix is then reduced to two smaller problems. The first is the evaluation of the selected optimality criterion for any particular tree, and the second is the search over the large space of trees for the particular tree that optimizes the selected criterion. In this section, the parsimony and likelihood criteria are discussed, and methods for computing the scores of individual trees are described. The problem of searching for optimal trees will be considered in the next section.

1.2.1 Parsimony

Parsimony, one of the most common methods for inferring phylogenies, is also one of the oldest, dating back to its introduction by Edwards and Cavalli-Sforza [11] in 1964 (see Chap. 10 in Felsenstein [20] for a nice account of the history of the field of phylogenetics). The parsimony method in phylogenetics is based on the general principle that simpler hypotheses should be preferred over more complex ones, where "simplicity" in the phylogenetic context is translated to mean the least amount of evolutionary change. Thus, trees that minimize the total amount of evolutionary change for a given data set are preferred, and the tree requiring the minimum number of evolutionary changes to explain the given data is called the most parsimonious or maximum parsimony (MP) tree.

Because the parsimony criterion is concerned with minimizing the amount of postulated evolutionary change, it can be applied to a variety of genetic data – essentially all that is required is a mechanism for quantifying "evolutionary change" in the observed data. The criterion can then be evaluated for any given tree by computing the amount of change required by that tree for the observed characters. To be more precise, consider a particular character, say x, and let x_i^h be the state of character h at node i in the tree, $1 \leq i \leq 2n-2$, where nodes 1 through n are external nodes corresponding to the tips of a rooted tree for which the character states are observed, and nodes $n+1$ through $2n-2$ are internal nodes whose character states must be inferred. Define $C(x_i^h, x_j^h)$ to be the cost of changing from the state for character h at node i to the state for character h at node j over the branch connecting nodes i and j. Note that it does not have to be the case that $C(x_i^h, x_j^h) = C(x_j^h, x_i^h)$, though equality is commonly assumed. The parsimony score of a tree, τ, under this criterion is then given by

$$S(\tau) = \sum_{h=1}^{N} \sum_{b=1}^{B} C(x_{b_1}^h, x_{b_2}^h), \tag{1.2}$$

where N is the number of characters in the data set, B is the number of branches in the tree, and b_1 and b_2 are the nodes at the ends of branch b, for which either the character state has been observed, or an optimal character state has been assigned. From (1.2), we see that the length of a tree is computed by summing lengths over all branches for a particular character in the data matrix, and then summing over all characters in the data matrix. Performing the calculation in this way necessarily assumes that changes among states on the branches occur independently once the states at all nodes are known, and also that changes across characters are independent. A weight, w_h, can be added in front of the cost term in (1.2) to allow for differential weighting of the characters in a data matrix.

The most commonly used cost function is one in which $C(x_i^h, x_j^h)$ is 1 if $x_i^h = x_j^h$ and 0 otherwise. This cost function counts the number of changes between character states in the tree and weights each type of change equally, i.e., any differences in state at the two ends of a branch increase the score of the tree by one, regardless of what those states are. This cost function can be applied to unordered multistate data, which can include molecular data (nucleotide and amino acid) as well as morphological data. This is generally referred to as Fitch parsimony. However, many variations of this cost function are possible. For example, for nucleotide data, it may be sensible to assign a lower cost to transitions than transversions, while for amino acid data, we might assign different costs for synonymous vs. nonsynonymous changes. For morphological data, we might specify an ordering in the data. For example, it is unlikely for beak size to change from "small" to "large" without first being "medium," and so an observed change from "small" to "large" might incur a cost equal to the sum of the costs of changing from "small" to "medium" and "medium" to "large." This is an example of what is known as Wagner parsimony [41, 18], for which costs are assigned to ordered multistate data in such a way that a change from one state to another incurs the sum of the costs of any intervening states.

To illustrate the computation of the parsimony length of a tree, consider the morphological data for the cephalopod example. We consider two trees representing alternative hypotheses concerning the placement of Vampyromorpha as described in Lindgren et al. (2004). These trees are shown in Fig. 1.1, and the observed states for character 38 in the data matrix are given at the tips of the trees (the complete data set is given in Lindgren et al. (2004)). Note that for this character, all species of Decabrachia have state 1, and thus the length of the clade containing the Decabrachia is 0. Thus, the Decabrachia clade has been collapsed into a single node in the trees in Fig. 1.1. To calculate the length of the trees, consider first the tree in Fig. 1.1a, and consider the node ancestral to *S. syrtensis* and *T. guntheri*. Since both *S. syrtensis* and *T. guntheri* have state 0, this ancestral node can also be assigned state 0, and this assignment requires no changes along the branches descending from it. Next, consider the node ancestral to *V. infernalis* and the ancestor of *S. syrtensis* and *T. guntheri*. Looking at this node's two descendants, we see that one of them (*V. infernalis*) has state 1 and the other (the ancestor of *S. syrtensis* and *T. guntheri*) has state 0. In this case, we could assign either state as a possible ancestral state. Note that selection of either state as the ancestral state will lead to one change along the tree, and thus we increase the length of the tree by one. Moving to the next most ancestral node, we see that the Decabrachia have state 1 and thus an assignment of state 1 to the ancestor of the Decabrachia and the clade containing *V. infernalis*, *S. syrtensis*, and *T. guntheri* is most parsimonious, and does not increase the length of the tree. Finally, the node at the root of the tree can be assigned state 1 without increasing the length of the tree. The overall length of this tree for this character is then 1.

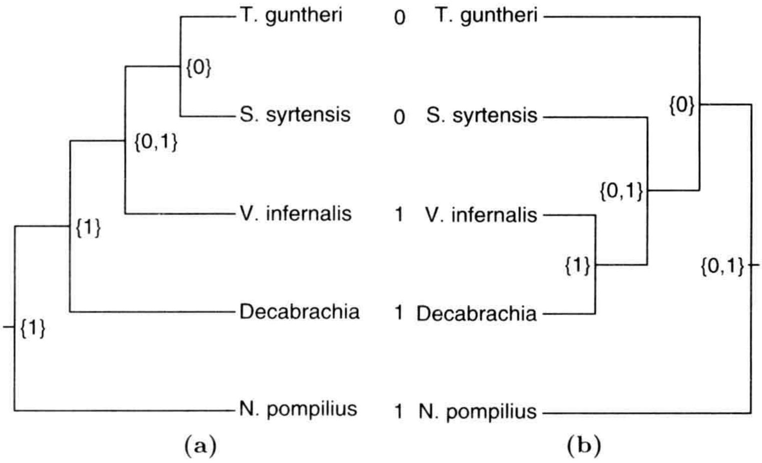


Fig. 1.1. Trees representing distinct hypothesized relationships among cephalopods, taken from Lindgren et al. (2004). The tree in (**a**) is the consensus of the nine MP topologies for the morphological data (each MP tree has length 107; the consensus tree has length 109). The tree in (**b**) is supported by the molecular data. Sets at the internal nodes of the tree are used to compute the score of the tree under the Fitch algorithm (see text for details)

This algorithm for computing the length of a tree, called the Fitch algorithm [26], can be expressed in more mathematical terms as follows. For each node in the tree, a set of character states will be assigned. The set at the tips of the tree contains a single state, the observed state at that tip. Then, for any node for which a character state set has been assigned for its two immediate descendants, the state set assigned to that node is the intersection of the state sets of its two immediate descendants if that intersection is nonempty; otherwise, it is the union of the state sets of the two immediate descendants. Whenever a union of state sets is required, the length of the tree is increased by one. The Fitch algorithm was developed specifically for unordered multistate characters, such as nucleotide and protein data, for which any state can change directly to any other state. Since changes in either direction are weighted equally under this method, a tree can be arbitrarily rooted with no change to its length, which allows one to root the tree at the most convenient location.

Comparing the trees in Fig. 1.1a,b, we see that for this character, the tree in Fig. 1.1a has length 1 and the tree in Fig. 1.1b has length 2, and so the tree in Fig. 1.1a is preferred for this character. Of the 45 parsimony informative characters in this data set, nine are informative for selecting between these two trees. Of these nine, eight favor tree (a) (characters 10, 38, 40, 45, 49, 57, 59, and 60) and one (character 6) favors tree (b). The result is that analysis of the morphological data favors placement of Vampyromorpha with the Octobrachia rather than the Decabrachia, which conflicts with the results obtained from the molecular data, as we will see in later sections. Figure 1.1a shows the consensus tree obtained from the nine MP trees (each MP tree has length 107; the consensus tree has length 109).

A second algorithm for computing the score of a tree under parsimony is the Sankoff algorithm. This algorithm works by assigning a function to each node of the tree which records, for each possible state, the minimum score for the subtree rooted by that node. We denote this function by $S_i^h(x)$, and define it to be the minimum score for the subtree rooted by node i assuming that node i has state x for character h. This value can be computed for any node for which this function has already been computed for its two immediate descendants using the following relationship

$$S_i^h(x) = \min_{x_j^h}\{C(x_i^h, x_j^h) + S_j^h(x_j^h)\} + \min_{x_k^h}\{C(x_i^h, x_k^h) + S_k^h(x_k^h)\}, \tag{1.3}$$

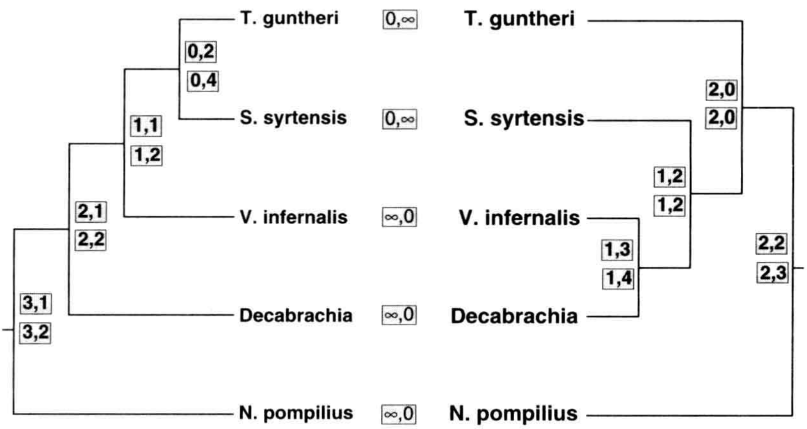
where j and k are the two nodes directly descending from node i. This equation is very intuitive. For example, consider the first term. This term corresponds to the branch descending from node i to node j. This branch contributes to the length of the subtree descending from node i in two ways: first, it contributes a length along the branch connecting nodes i and j; second, it contributes a length due to the subtree descending from j, as recorded by the S function for node j. There is then a similar contribution from the other branch descending from node i, denoted by k here. Taking the minimum over all possible assignments of states to the nodes j and k will give the minimum at node i, given that it has state x.

This algorithm is applied successively to the nodes of the tree in a postorder traversal (see Felsenstein ([20], p. 587)). The value of the S function at the tips of the tree is determined by setting $S_m^h(x) = 0$ if tip m has state x for character h, and $S_m^h(x) = \infty$ otherwise. The minimum length of the entire tree is then found by selecting the minimum value of the S function at the root of the tree. Denoting the root node by r, the parsimony score of the tree is

$$S(\tau) = \sum_{h=1}^{N} \min_x S_r^h(x). \tag{1.4}$$

Figure 1.2 gives an example of the computation for the morphological data for the cephalopod example, with both the simple cost matrix used in the explanation of the Fitch algorithm, and a modified cost matrix that results in a different conclusion concerning which tree (of the two) is the most parsimonious.

We note that both the Fitch and Sankoff algorithms are dynamic programming algorithms, since they reduce the problem of computing the score to subproblems, which can be optimally solved in such a way that it can be proved that they lead to the overall optimal solution. Felsenstein ([20]; p. 16) discusses the connection between the two methods. The Sankoff algorithm is



(a) For the first cost function, $S = 1$, while for the second cost function, $S = 2$.

(b) For both cost functions, $S = 2$.

Fig. 1.2. Trees representing distinct hypothesized relationships among cephalopods, taken from Lindgren et al. (2004). The tree in (**a**) is the consensus of the nine MP topologies for the morphological data. The tree in (**b**) is supported by the molecular data. The colored boxes at the nodes of the tree represent the $S()$ function used to compute the length of tree under the Sankoff algorithm for two different cost functions. The upper (*blue*) boxes at each node correspond to the same cost function as was used to illustrate the Fitch algorithm: $C(0,0) = C(1,1) = 0$; and $C(0,1) = C(1,0) = 1$. The lower boxes (*yellow*) correspond to a cost function that penalizes more for one particular change: $C(0,0) = C(1,1) = 0; C(0,1) = 1$; and $C(1,0) = 2$. For the first cost function, the tree in (**a**) is preferred, while for the second cost function, the scores of the two trees are equivalent

more general, in that it allows the use of *any* cost function, while the Fitch algorithm is confined to the setting where all changes are weighted equally. We also note that while both algorithms specify a sum over characters to compute the total score for the tree, the computation can be simplified for both algorithms by computing the scores for only unique sites. For example, any character for which all taxa have the same state will require no changes on every tree. Additionally, under Fitch parsimony, any character for which all taxa except one have the same state will require exactly one change on any tree. Characters of this nature are generally said to not be *phylogenetically informative*, since they do not prefer any tree over any other in the parsimony setting. Therefore, no computations need be performed on these character patterns. However, these character patterns do contribute to estimation in other settings, as will be seen for likelihood in the following section. For a particular cost function, there may also be other classes of characters for which the score will be identical, and therefore needs to be computed only once and then multiplied by the number of characters observed in that class. An example will be given below.

此为试读,需要完整PDF请访问: www.ertongbook.com