Parsimony and the problem of inapplicables in sequence data.

De Laet, J. 2005.

Pp. 81-116 in
Albert, V.A. (ed.) Parsimony, Phylogeny and Genomics.
Oxford University Press, ISBN 0-19-856493-7.

CHAPTER 6

Parsimony and the problem of inapplicables in sequence data

Jan E. De Laet

''I don't know what you mean by 'glory,'' Alice said. Humpty Dumpty smiled contemptuously. 'Of course you don't–till I tell you. I meant 'there's a nice knock-down argument for you!'' 'But 'glory' doesn't mean 'a nice knock-down argument,'' Alice objected. 'When I use a word,' Humpty Dumpty said in rather a scornful tone, 'it means just what I choose it to mean–neither more nor less.''

(Caroll 1872, chapter VI)

6.1 Introduction

About 10 years ago, Maddison (1993; see also Platnick *et al*. 1991) drew attention to problems that can arise in parsimony analyses when data sets contain characters that are not applicable across all terminals. Examples of such characters are tail color when some terminals lack tails, or positions in DNA sequences in which gaps are present. Maddison (1993) examined various ways of coding such characters for various parsimony algorithms and concluded that no general solution was available. Since then, the problem of inapplicables has been rediscussed repeatedly (e.g. Lee and Bryant 1999; Strong and Lipscomb 1999; Seitz *et al*. 2000), but Maddison's conclusion still holds.

Farris (1983), focusing on regular single-column characters as classically used in phylogenetic analysis, characterized parsimony as a method that maximizes explanatory power in the sense that most-parsimonious trees are best able to explain observed similarities among organisms by inheritance and common ancestry. This led De Laet (1997; see also De Laet and Smets 1998) to formulate parsimony analysis as two-item analysis. In this view, parsimony maximizes the number of observed pairwise similarities that can be explained as identical by virtue of common descent, subject to two methodological constraints: the same evidence should not be taken into account multiple times, and the overall explanation must be free of internal contradictions.

Here, I examine how this formulation can be used to deal with the problem of inapplicables. More specifically, I deal with the problem of inapplicables in sequence data, a harder and more general problem than most cases of inapplicability that Maddison (1993) had in mind. The review of parsimony analysis in the first section provides the basis for discussing the analysis of sequence data in the second section. The basic idea of the whole chapter is to explore the ramifications of the conceptual framework of Farris (1983) beyond the realm of single-column characters. This was in part prompted by the double observation that several authors seem to be using isolated elements of that paradigm when discussing methods for sequence analysis (see, e.g., Frost *et al*. 2001; Simmons 2004), while, at the same time, no coherent discussion of those ideas as applied to sequence data is available.

6.2 Parsimony analysis as two-item analysis

Some notes on terminology are appropriate first. Take a simple term such as 'autapomorphy'.

Originally, autapomorphies were defined as 'apomorphous features characteristic for a particular monophyletic group (present only in it)' (Hennig 1966, p. 90). In addition to this original meaning, a more restrictive usage that reserves the term for 'novelties that are coded as unique in a data set' (Kluge 1989, p. 9) is widespread.

	Characters									
	c1	*c2*	*c3*	*c4*	*c5*	*c6*	*c7*	*c8*	*c9*	*c10*
Terminals										
out1	0	0	0	1	0	0	0	1	0	0
out2	0	0	0	1	0	0	0	0	0	0
A	1	0	0	1	0	2	0	0	1	1
B	1	1	1	0	0	1	1	0	1	0
C	1	1	1	0	0	1	0	0	1	0
D	1	1	1	1	1	2	0	0	1	1
E	1	1	1	1	1	2	0	0	0	1
F	1	1	1	1	1	2	0	1	0	1

Figure 6.1 A data set with 10 unordered characters for eight terminals. Terminals *out1* and *out2* are interpreted as outgroups.

Consider the data set of Fig. 6.1 and its most-parsimonious tree (*out1 out2* (*A* ((*B C*) (*D* (*E F*))))) (see Fig. 6.2). Under Hennig's original definition, the first seven characters all provide autapomorphies. As an example, character *c4* has apomorphous state *0* for monophyletic group (*B C*), and that state does not occur outside that clade. Under the more restrictive definition only character *c7* is autapomorphic. Obviously, questions as to whether autapomorphies should be taken into account or not when calculating the consistency index of a data set on a tree (e.g. Yeates 1992) take an entirely different meaning depending on the way in which the term 'autapomorphy' is used.

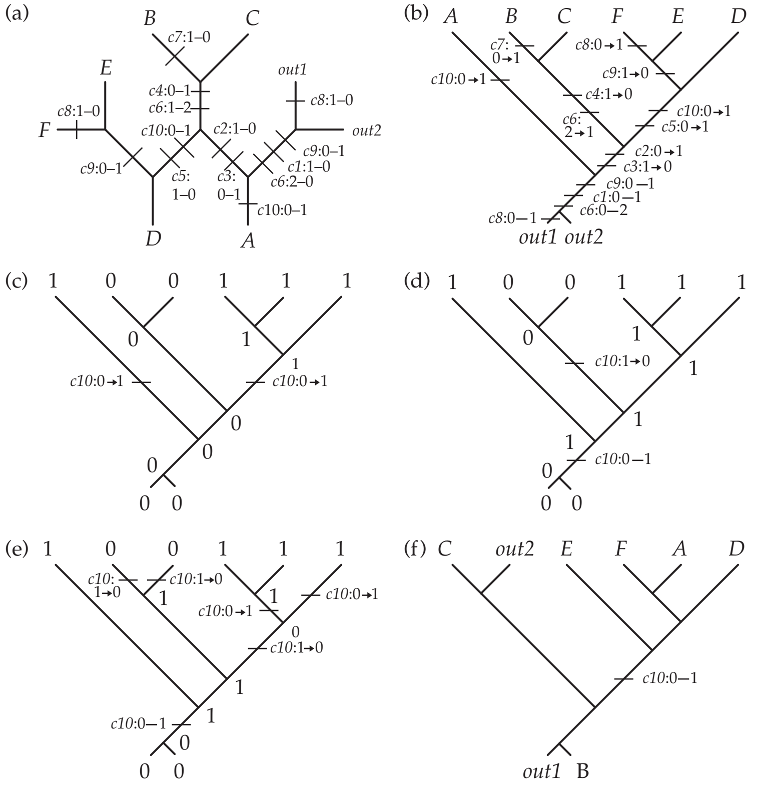

Figure 6.2 Parsimony analysis of the data of Fig. 6.1. (a) The most-parsimonious explanation of the data requires 14 steps. (b) To come to hypotheses of synapomorphy and monophyly in the ingroup, the ingroup is rooted using the branch that leads to the outgroups (note that this procedure does not imply such hypotheses outside the ingroup). (c, d) Two alternative optimal explanations of character *c10* on the most-parsimonious tree. (e) A suboptimal explanation of character *c10* on the most parsimonious tree. (f) An optimal explanation of character *c10* on a suboptimal tree.

Paraphrasing Farris (1983, p. 8), I share Humpty Dumpty's disdain for arguing definitions as such. Therefore I shall not discuss and evaluate the pros and cons of various possible meanings of the terms that I employ, nor indicate alternative terms with identical or similar meanings. But as the above example shows, it is important to make intended meanings clear, so in this section I shall explicitly point out my usages of terms.

At the same time, this process will provide an interlocked set of concepts that will allow a clear discussion of parsimony and inapplicables in the next section, and help to distinguish terminological issues from more substantial argument. To preempt any objection as should the conclusions hinge on major redefinitions of familiar terms, I shall indicate how my usages are rooted in existing literature. This, however, should not be taken to imply that these usages are always strictly in line with those references: whenever some existing, term is close enough, in spirit, to intended use (as would, e.g. Kluge's use of Hennig's autapomorphy above) I shall adopt existing terminology rather than propose a new term.

6.2.1 Characters and character analysis

Conceptually, a cladistic analysis consists of two main activities (see, e.g., Rieppel 1988; de Pinna 1991; Rieppel and Kearney 2002). The first comprises empirical observation, leading to delimitation of characters and character states, and to a data set in which those characters are scored for the terminals in the analysis. This is the activity of perceiving similarity and coding it into characters and data sets, to which I shall refer as *character analysis* (Kluge and Farris 1969, p. 9–10; see also Rieppel and Kearney 2002, p. 60). The second activity takes data sets as input, identifies their most-parsimonious hierarchic arrangment(s), and uses the resulting cladogram(s) as a basis for phylogenetic inference. I shall refer to this as *parsimony analysis* (Farris 1983, p. 10–12; see also later).

Character analysis and parsimony analysis stand in a continuous relationship of reciprocal illumination, at different levels (e.g. Rieppel 2003, p. 182; see also Hennig 1950, p. 26). As an example, the selection of terminals that will be included in a data set is in part guided by existing phylogenetic hypotheses. Likewise, empirical work that results in new characters that are added to data sets can lead to cladograms with new or refined hypotheses of phylogenetic relationships. These, in turn, can point to characters that are highly incongruent with the general pattern and that may therefore be worth additional scrutiny. If an empirical basis can be found for a reinterpretation of such characters or their states, the data set can be adapted accordingly (see, e.g., Farris 1983, p.10).

At a given point in this process of continuous refinement, consider an individual character such as *c4* in the data set of Fig. 6.1. From the point of view of character analysis this *character* is a statement about a feature that comes in two states, coded *0* and *1*, such that state *0* is observed in terminals *B* and *C* and state 1 in all other terminals. Theoretically, such a character expresses the hypothesis that the observed feature carries evidence on the genealogical relationships among the taxa that are involved. This directly limits characters and character states for phylogenetic analysis to features that are inheritable. A thought-provoking discussion of this seemingly trivial observation can be found in Freudenstein *et al*. (2003).

Beyond this, however, little more specific can be said other than that a character state as observed in different terminals 'must be sufficiently similar to be called the same [. . .] at some level of taxonomic generality' (Kluge 1997a, p. 89; the quote refers to derived states but the statement is valid in general), an observation that also holds for the character as a whole (see, e.g., Platnick 1979, p. 542; Jenner 2004, p. 301). For morphological and anatomical features, the criteria of composition, conjunction, ontogeny, and topography provide perspectives that can serve to evaluate if such sufficiency holds in particular cases (Kluge 1997a). Of those, topography or topological relationships are often considered to be the fundamental criterion (e.g. Rieppel 1988; de Pinna 1991, Hennig 1966, pp. 93–94; see also Remane 1952, pp. 31–66).

As discussed extensively by Rieppel and Kearney (2002, in the context of anatomy; see also Jenner 2004), care must be taken to give similarity statements as expressed in characters an observational basis. In order to do so one has to rely, however, unavoidably on background knowledge, and there is in principle no limit to the degree of background knowledge that can be incorporated in a character (Rieppel and Kearney 2002, p. 265). So even in this specific and restricted context of erecting character hypotheses for cladistic analysis, the concept of similarity unavoidably retains some elusiveness. This notwithstanding, similarity assessments as expressed in characters and their states, in the theoretical framework as just dicussed, are the empirical basis on which further phylogenetic inference is built.

6.2.2 Single-character phylogenetic inference

If no other comparative data were available for the terminals that are involved, a character such as *c4* would constitute a data set on its own. It is a useful exercise to subject such a minimal data set to parsimony analysis. Within the constraint of terminal sampling, this leads to the following inferences: (1) the feature arose in a common ancestor of these terminals, from which they inherited it; (2) differentiation into two states ocurred at a later stage; (3) for each state, the terminals with that state are only connected through ancestors that have that same state. These inferences do not yet include a polarity statement for which state is considered apomorphic and which plesiomorphic.

The apomorphy/plesiomorphy pair of terms is defined as follows: for a given evolutionary transformation, the condition or state from which the transformation started is *plesiomorphic* or *primitive* and the condition after the transformation *apomorphic* or *derived* (Hennig 1966, p. 89). As discussed by Hennig (1966, p. 93), coming to an hypothesis of features that are involved in such a transformation on the one hand and deciding on the evolutionary direction of such a transformation on the other are entirely different questions. The inclusion of outgroups in data sets is arguably the most general and least assumption-laden way to address the latter question.

Roots and outgroups

In general, when studying the phylogenetic relationships among a group of terminals, one assumes that these are part of a *monophyletic group* at some level of inclusiveness, meaning that they share a common ancestor that they do not share with terminals outside that group (Hennig 1966, 73–74; see Farris 1991 for a review of this and related terms). The terminals that are assumed to be part of the monophyletic group are called *ingroup terminals* and are collectively referred to as the *ingroup*. Terminals outside that group are called *outgroup terminals* or *outgroups* for short.

When outgroups are included in a data set, they can be used to root the ingroup after the globally most-parsimonious arrangements of the data have been identified (Farris 1972, p. 657; see Figs 6.2a and 6.2b for an example). In the ingroup, hypotheses of relative apomorphy and plesiomorphy and of the direction of transformations then directly follow (Farris 1982a; see Figs 6.2c and 6.2d for some examples). This is the procedure that is now almost universally used to root ingroups and polarize characters, and it is mostly referred to as the *outgroup method* or the *outgroup criterion* (see, e.g., Farris 1979, p. 511). Confusingly, these and similar labels were also used in a series of papers in the 1980s for a series of methods of prior character polarization that are fundamentally different and mostly no longer in use. A historical account and a discussion of these methods can be found in Nixon and Carpenter (1993). The precise way in which hypotheses on character polarity come about does not affect the argumentation in this paper, so without loss of generality the discussion is restricted to outgroups.

In a data set that has only one character, as above, the general use of outgroups as just described becomes simplified because the best tree for the data set coincides with the structure of its single character. In the above example, the outgroup hypothesis could be the assumption that terminals *A* through *F* (the ingroup) share a most recent common ancestor that is not shared with terminals *out1* and *out2* (the outgroups). Observing that state 1 of character *c4* is present in the outgroups as well as in the ingroup, it follows that state 1 is plesiomorphic in the ingroup; that state 0 is apomorphic in that same group; and that (*B C*) is a monophyletic subgroup of the ingroup.

Outgroups do not always lead to such unambiguous single-character inferences. An example is character *c6*, where (*A D E F*) and (*B C*) could both be monophyletic; or, alternatively, either could be paraphyletic with the other monophyletically nested in it. In addition, contradictions can arise between a character hypothesis and the outgroup hypothesis, even with binary characters. An example is character *c8*: the two following statements, derived from the character, contradict the outgroup hypothesis: terminals *out1* and *F* are only connected through ancestors that have state 1; the other terminals are only connected through ancestors that have state 0. Such cases are mostly but not necessarily interpreted to mean that the hypothesis of ingroup monophyly is incorrect. In general, nothing more can be said other than that the data do not support the prior assumption of ingroup monophyly (Farris 1972, p. 657), an observation that is also consistent with the alternative interpretation that the data are wrong. Neither issue addressed in this paragraph affects the argumentation of this paper.

Premises

Obviously, the above conclusion of monophyly for (*B C*) is conditional: it depends on the correctness of the outgroup hypothesis, on the correctness of the similarity assessments that led to character *c4* and its coded states, and on the correctness of several other, hidden, assumptions that remained unexpressed (such as absence of reticulate evolution). So, it would be more precise to say that (*B C*) is a putative monophyletic group, or a presumed monophyletic group, or that *B* and *C* are hypothesized to be monophyletic, each time conditional on the premises stated above (see Farris 1983, p. 13 for a similar use of the term 'putative'). Below, I shall use such verbose formulations only when confusion could arise otherwise, or when I wish to stress the difference between hypothesis or inference on the one hand and true historical account on the other. For the latter I shall then use the convenient adjective 'true', following existing practice (see, e.g., Farris 1983, p. 12), while observing that the philosophical problems that surround the notion of truth (see, e.g., Boyd 1991) do not affect this usage. The same applies to some other terms that I already have used: outgroup, apomorphy, and plesiomorphy are defined in terms of phylogenetic history but are often used to refer to just a hypothesis about that history.

Hennig (1966, p. 89) introduced the terms *symplesiomorphy* and *synapomorphy* to decribe the presence of plesiomorphies and apomorphies among terminals. As above, these terms are defined with respect to true evolutionary history, but are often used to refer to inferences as well. Such context-dependent shifts in meaning of these and similar terms are widespread in the literature, Hennig (1966) being a prime example. Related to this, when considering a transformation series such as $a \rightarrow a'$, Hennig (1966, pp. 88–89) sometimes referred to a and a' as 'character conditions,' sometimes as 'special characters' and sometimes even just as 'characters.' Combined with context-dependent meanings of terms, such use of different terms for the same thing, with meanings that often differ from current usage, can make it hard to understand Hennig's writings. This is even more problematic because Hennig used an argumentation scheme to order and polarize characters that is very different from current practice. In the above example, Hennig referred to a and a' as characters 'in the sense that they distinguish their bearers from one another' (Hennig 1966, p. 89). At the level of character analysis they are, in current usage, just character states.

When used conditionally, the precise meaning of terms such as synapomorphy and plesiomorphy in particular cases can drastically change according to the exact conditionals that are used or implied. Consider, for example, isolated character *c9* and the outgroup hypothesis. In that case the presence of state 1 in terminals *A*, *B*, *C*, and *D* is a (putative) synapomorphy compared to the presence of state 0 in terminals *out1*, *out2*, *E*, and *F*, which is a (putative) plesiomorphy. On the other hand, when considering the whole data set of Fig. 6.1 and its most-parsimonous tree (Fig. 6.2b), the presence of the same character state 1 in the same terminals *A*, *B*, *C*, and *D* is now a (putative) symplesiomorphy compared to the presence of state 0 in terminals *E* and *F*, which has become a (putative) synapomorphy. The presence of state 0 in the outgroups remains a (putative) symplesiomorphy. More interestingly, the presence of apomorphic state 1 in its original form (terminals *A*, *B*, *C*, and *D*) and in its more derived form (terminals *E* and *F*) is now a putative synapomorphy for terminals *A*–*F*.

6.2.3 Homology, the Hennig–Farris auxiliary principle, and parsimony analysis

A crucial assumption in the above interpretation of a single character is *Hennig's auxiliary principle*, stating 'that the presence of apomorphous characters in different species . . . is always reason for suspecting kinship [i.e. that the species belong to a monophyletic group], and that their origin by convergence should not be assumed a priori' (Hennig 1966, p. 121; square brackets present in original). In this quote, the term 'character' refers to a 'special character' (Hennig 1966, p. 89), which is a character state as used in this chapter, whereas an apomorphous (special) character refers to a special character that 'can certainly or with reasonable probability be interpreted as apomorphous' (Hennig 1966, p.121), i.e. an hypothesis of apomorphy or a putative apomorphy; monophyly is used in its true historical meaning.

Without this principle, one could equally well assume that, for example, state 1 of character *c5* of Fig. 6.1 arose multiple times. As an example, on the most-parsimonous tree (Fig. 6.2b) state 1 could have arisen a first time in the branch that leads up to terminal *D*, and a second time in a common ancestor of *E* and *F* that is not a common ancestor of *D*. Under this interpretation, the shared presence of 1 in *E* and *F* would be interpreted as evidence for monophyly of clade (*E F*), to the specific exclusion of terminal *D*, even if *D* has the same state.

However, given that the delimitation of character *c5* is grounded in empirical observation, this is not a very plausible interpretation of the character. Indeed, if any empirical evidence were available that state 1 as present in terminal *D* is not sufficiently similar to state 1 as found in terminals *E* and *F* to be called the same at some level of generality, these terminals would not have been assigned the same numeric state code to begin with. Since this was not the case, preferring the second interpretation over the first amounts to discarding some of the evidence that bears on the problem at hand (viz. the perceived similarity between terminal *D* on the one hand and terminals *E* and *F* on the other. The remaining evidence (viz. the perceived similarity between *E* and *F*) then supports monophyly of *E* and *F* to the exclusion of *D*.

Homology should be presumed in the absence of evidence to the contrary

Hennig's formulation of his auxiliary principle, quoted earlier, is logically inconsistent because it can lead to internal contradictions: if the presence of presumed apomorphies is always to be a reason for suspecting true monophyly (first part of the principle), then it is not simply sufficient that multiple, convergent, origins of that state should not be assumed *a priori* (second part). This would still leave open the possibility that some terminals with the presumed plesiomorphic state obtained that state through a reversal. In that case, the group of all terminals with the presumed apomorphic state would no longer be truely monophyletic, which contradicts the first part. So that first part by logical necessity requires an additional statement that the origin of presumed plesiomorphies should not *a priori* be interpreted as reversals (for characters with more than two states, a similar statement is required for each state). As an example, without this addition a character such as *c5* could be taken as evidence for, e.g., a monophyletic group (*A D E F*) because it is not precluded that state 0 in terminal *A* arose as a reversal within that clade. In this interpretation, state 0 as present in terminal *A* would be derived relative to state 1 as present in terminals *D*, *E*, and *F*.

Such additional statements are implicit in Farris' (1983, p. 8) formulation of Hennig's auxiliary principle: 'homology should be presumed in absence of evidence to the contrary', where *homology* refers to similarities among organisms that have arisen historically through inheritance from a common ancestor, irrespective of these similarities being apomorphic or plesiomorphic. More explicit discussions of the necessity, in parsimony analysis, of explaining plesiomorphic similarities as due to common descent can be found in Farris *et al*. (1995, p. 215) and

Farris (1997, pp. 132–133). I shall therefore refer to the auxiliary criterion in its logically consistent form as the *Hennig–Farris auxiliary principle*.

When, as above, the Hennig–Farris auxiliary principle is applied to single–character data sets, it can be interpreted as a condition that makes the apomorphic state by necessity mark a true monophyletic group: the state arose only once and never reverted. That group will be present on any tree that requires only a single origin for that state, which is in line with Farris' (1983, p. 12) observation that grouping by true synapomorphy would have to behave exactly as parsimony, in the sense that it would lead to preference for the tree(s) on which no homoplasy is present (*homoplasy* being a point of similarity among organsims that cannot be explained by inheritance and common descent on a particular tree; Farris 1983, p. 18; see also below). These are, by definition, the shortest trees possible, so they are also most parsimonious trees.

Parsimony and the Hennig–Farris auxiliary principle

In practice, however, one is constrained to work with actual observable traits of organisms rather than with true historical synapomorphies. Character codings of such traits seldom if ever capture all true evolutionary transformations, let alone their order, as exemplified by the presence of homoplasy in all but the smallest and simplest data sets (note that absence of homoplasy in such data sets would hardly justify the conclusion that all relevant transformations have been captured—absence of evidence is not evidence of absence). This led Farris (1983, p. 17–19; see also Farris and Kluge 1986, p. 300; Farris 1986, pp. 15–16) to a general characterization of parsimony analysis in terms of a methodological principle that is fundamental to science in general: maximization of explanatory power or conformity between observation and theory. More specifically, the observations are the similarity statements as coded in characters, and the theory is that these similarities have arisen through inheritance and common descent. Most-parsimonious cladograms are then preferred because they are the trees on which the greatest amount of such observed points of similarity among organisms can be explained by inheritance and common descent (contra Grant and Kluge 2004, p. 29). As such they provide the best explanation of the observations on account of the theory.

Note that, at this level of analysis, characters and their states can indeed be treated as simple observations, even if, as discussed above, they are complex theories or hypotheses on their own. Likewise, little confusion arises if the presence of the same character state of a given character in two terminals is simply called an *observed point of similarity* between those two terminals. Such usages of these terms can be found, for example, throughout Farris (1983).

Similarities as coded in characters can very well be true homoplasies rather than true homologies. Likewise, it cannot be ruled out that character similarities that can be explained as homologies on most-parsimonious cladograms are true homoplasies instead, even when using single-character data sets as above. Combined with the observation that parsimony minimizes putative homoplasy, such observations are sometimes taken to mean that it is an assumption of parsimony analysis that homoplasy is rare in evolutionary history. However, even if rarity of homoplasy may be a sufficient condition to prefer most-parsimonious trees (see, e.g., Felsenstein 1981), it is definitely not a necessary condition.

Consider a data set for terminals *out*, *A*, *B*, and *C* where 10 characters support clade (*B C*) and just one character supports clade (*A C*) (this example and discussion is based on Farris 1983, pp. 13–14, see also p. 12, pp. 18–19). If clade (*A C*) is genealogically correct, then the 10 characters that support (*B C*) are (true) homoplasies; if, on the other hand, clade (*B C*) is genealogically correct, then the single character that supports (*A C*) is a (true) homoplasy. These simple observations point out an interesting asymmetry in the relationship between characters and genealogies: a given genealogy implies that characters that contradict this genealogy are homoplasious but requires nothing concerning characters that do not contradict the genealogy. Now assume that true homoplasy is so abundant that only one out of those 11 characters has escaped its effects. Under the assumption that this one character can equally well be any character in the data set, a simple statistical argument leads to preference for clade (*B C*): the probability that this single historically correct character supports this clade is 10 times higher than the probability that it supports (*A C*). Thus it is seen that even under extremely high levels of homoplasy most-parsimonious trees can still be the best phylogenetic hypotheses one can make on the basis of the available data, even if some of the putative homologies may be true homoplasies instead.

The underlying assumption of the above conclusion is best stated in the negative: absence of any assumption about the distribution of homoplasies in data sets. In a statistical framework, this can be understood as the use of an uninformative prior. Obviously, one can postulate distributions of homoplasy such that the most-parsimonious trees will no longer be the best bets. Such distributions are typically derived from stochastic models of sequence evolution (see, e.g., Felsenstein 1978a; Huelsenbeck and Lander 2003). The mere fact, however, that such distributions can be postulated does not by itself invalidate parsimony analysis as a method to analyze empirical data. Indeed, such a conclusion would crucially hinge on the realism or plausibility of the underlying stochastic models (and not on their simplicity, as Huelsenbeck and Lander 2003 seem to suggest). Farris (1983, pp. 14–17, p. 12; see also Farris 1999) amply discussed these issues and found the models that were in use at that time greatly lacking in realism. Stochastic models of sequence evolution have dramatically increased in complexity since then (see Felsenstein 2004 for a review), but they still seem mostly inadequate to model even small-sized real data sets (D. Pol, personal communication). Therefore, Farris' discussion and conclusions remain as valid and to the point as they were more than 20 years ago.

Considering all this, the Hennig–Farris auxiliary principle can be phrased as the following rule for erecting character hypotheses and interpreting their optimizations on trees: 'features that on the basis of empirical evidence are deemed sufficiently similar to be called the same at some level of generality should be treated as putative homologues in phylogenetic analysis (even if they may be true homoplasies instead).' In combination with the principle of maximizing explanatory power, this makes similarity-based statements of putative homology the centerpiece of phylogenetic inference: most parsimonious trees are trees on which the greatest amount of putative homology statements that return from character analysis *can* be explained as due to inheritance and common descent, and such trees are the best available phylogenetic hypotheses for the terminals at hand, whether or not the individual similarity statements or their explanations are historically correct.

As just discussed, the premises under which this holds are best stated in the negative: complete nonreliance on specific premises regarding correlations of evolutionary rates within and across characters and lineages. As such, parsimony analysis can be considered the most general method for phylogenetic analysis that is available. Tuffley and Steel (1997; see also Steel and Penny 2000) and Goloboff (2003) have examined similar but less extreme positions of agnosticism with respect to the details of evolutionary processes, using stochastic modeling. In both cases the most-parsimonious tree(s) are the best phylogenetic hypotheses, reinforcing the above conclusion.

6.2.4 Quantifying and maximizing homology

Given a tree and a data set such as in Fig. 6.1, Farris (1983) did not directly quantify the amount of points of similarity that can be explained by common descent and inheritance on that tree. Instead he used, as a relative measure, the minimum number of independent statements of homoplasy that are required on that tree. This works because an instance of homoplasy is present on a tree whenever a point of similarity as expressed in a character cannot be explained as homology on that tree (Farris 1983, p. 18).

So, when comparing two trees, the tree with the lower level of homoplasy will have the greater amount of similarity that can be explained as homology, and hence the greater power to explain the data on account of the theory. In practice, most parsimony programs calculate the minimum number of steps that are required, which, for a given character, differs from the minimum number of independent statements of homoplasy by a constant factor. As a result, the same ranking of trees is obtained. Several points are worth elaborating here.

Inner-node state assignments and the requirement of internal consistency

First, whether or not a particular pairwise similarity as coded in a character can be explained as a homology on a particular tree does not just depend on the structure of the tree and on the state distribution of the character that is involved, but also on assumptions that are made about the character states that are present at the internal nodes of the tree.

Take character *c10* of the data set of Fig. 6.1 and the most-parsimonious tree for that data set (Fig. 6.2b). Representing a pairwise similarity that is expressed as the presence of a same state *i* of a character in two terminals *X* and *Y* as $S_i(X\ Y)$, or, equivalently, $S_i(Y\ X)$, the similarity among terminals *A* and *D* as coded in *c10* is $S_1(A\ D)$. With inner node state assignments as in Figs. 6.2c or 6.2e, this pairwise similarity cannot be explained as a homology because independent derivations of state 1 from state 0 are involved. On the other hand, with state assignments as in Fig. 6.2d, that same similarity can be explained as a homology. Similarly, $S_0(out1\ B)$ can be explained as a homology in Fig. 6.2c but not in Figs. 6.2d and 6.2e. In general, a pairwise similarity $S_i(X\ Y)$ can be explained as a homology on a tree when all nodes that connect *X* and *Y* have been assigned that same state *i*; in that case, the statement is said to be *accomodated* on the tree. In all other cases, it is a homoplasy, and the statement is not accomodated (only cases in which unique states are assigned to inner nodes are considered in this paper; polymorphic inner nodes, as in Farris (1978a) or in Felsenstein (1979), are left undiscussed).

The connection between the explanation of a character and assignments of states to inner nodes can be seen as a methodological constraint that ensures that the set of all homology statements that can be derived from a tree and a character state distribution is free from internal contradictions (De Laet and Smets 1998, pp. 374–376). Or, put positively, it ensures that the overall explanation is logically possible or consistent. This, in turn, makes the explanation of the character on the tree logically capable of phylogenetic interpretation (Farris *et al*. 2001b). For example, on this tree one can explain either the similarity between *A* and *D* (e.g. Fig. 6.2d) or the similarity between *out1* and *B* as a homology (e.g. Fig. 6.2c); one cannot possibly, however, simultaneously explain both similarities as homologies because they are mutually exclusive. This logical requirement of non-contradiction is also met in maximum likelihood methods that integrate over all possible sets of inner-node state assignments, such as that of Felsenstein (1981). It is not met in quartet and triplet methods (De Laet and Smets 1998). Pairwise similarity statements that can simultaneously be explained as homology on a given tree will be referred to as (mutually) *compatible statements*.

When the terminals of a tree are labeled with the observed states of a particular character and the inner nodes have been assigned character states as well, the tree can be cut into a number of parts in which all nodes have the same state, and such that neighboring parts have different states. I shall refer to such parts as *regions*. There is a straightforward connection between number of regions and number of steps: any boundary between two regions implies a step, so the number of steps is one less than the number of regions. By definition, all similarities within a region can be explained as homologies, while similarities across regions are homoplastic. Because these regions are non-overlapping and because homologies do not cross the borders of such regions, the problem of quantifying the amount of similarity of the character that can be explained as homology on the tree can be broken down easily into the smaller problem of determining the amount of homology in such a region. For the same reason, the different states of a character can be treated independently under those conditions.

Independence and the units of empirical content of comparative data sets

A second issue is logical independence of pairwise homology (and homoplasy) statements within characters (Farris 1983, pp. 19–20, 21–22; De Laet and Smets 1998, pp. 369–374; this is different from logical dependence *between* characters, as discussed, e.g., in Wilkinson 1995, pp. 297–298). Consider state 1 of character *c10* as it returns from character analysis. At that point, all its six pairwise similarity statements can be interpreted as homologies: $S_1(A\ D)$, $S_1(A\ E)$, $S_1(A\ F)$, $S_1(D\ E)$, $S_1(D\ F)$, and $S_1(E\ F)$. Not all of these are independent though: if, e.g., $S_1(A\ D)$ and $S_1(A\ E)$ can be interpreted as homologies, then, by necessity, $S_1(D\ E)$ can be interpreted as a homology as well. In general, if n_i terminals have the same character state for a given character, there are $n_i * (n_i - 1)/2$ different pairwise similarity statements that can be made, but no more than $n_i - 1$ of those can be independent. Adding statements beyond this number will introduce redundancy in the description of the data. This maximum number of independent pairwise similarity statements is at the same time the minimum number of statements that must be considered to deduce the complete set: when removing statements from a largest set of independent statements, there is no longer sufficient information to generate all data.

Non-redundant descriptions. I shall call such maximal sets of independent pairwise similarity statements *smallest generating sets*. The exact identity of the members of such sets does not matter, the important points are completeness and absence of logical dependencies. As an example, $\{S_1(A\ D), S_1(A\ E), S_1(A\ F)\}$ and $\{S_1(A\ D), S_1(D\ E), S_1(E\ F)\}$ are two different smallest generating sets for state 1 of character *c10*; $\{S_1(A\ D), S_1(A\ E), S_1(A\ F), S_1(E\ D)\}$ is a generating set, but not a smallest one because not all of its elements are independent. Next consider how the pairwise similarities in a character state can be explained on a particular tree with a particular set of inner-node state assignments, such as, for example, in Fig. 6.2c. There are two regions that have character state 1: isolated node A and subtree (D (E F)). All similarities within a region are homologies and all similarities across regions homoplasies, so $S_1(D\ E)$, $S_1(D\ F)$, and $S_1(E\ F)$ are homologies, while $S_1(A\ D)$, $S_1(A\ E)$, and $S_1(A\ F)$ are homoplastic.

A non-redundant description of this can be determined as follows. For each region that is involved, establish a smallest generating set (in general, a region with j terminals will have smallest generating sets of cardinality $j - 1$). These sets non-redundantly describe the homologies of the character state on the tree, and the total number of independent statements that are accomodated is the total number of statements in these sets. Then pool these generating sets and augment the resulting set to obtain a smallest generating set for all similarities in the character state, without reference to a tree. The added statements form a maximal set of independent pairwise similarity statements that are not accomodated. This procedure establishes that the number of independent accomodated homologies and homoplasies for a given state add up to a number that is tree-independent. As a result, minimizing the number of independent statements of pairwise homoplasy in a character state and maximizing the number of independent statements of pairwise homology in that same state are equivalent problems indeed. Because independent homologies can be counted one region at a time, this remains true when summing over all states in a character, and/or over all characters in a data set.

In this example, the first region (isolated node A) has no similarities and therefore an empty smallest generating set; $\{S_1(D\ E), S_1(E\ F)\}$ is a smallest generating set for the second region. Adding, for example, homoplastic statement $S_1(A\ E)$ is sufficient to fully describe the character state and its explanation on the given tree. As an example, given that $S_1(D\ E)$ is accomodated and that $S_1(A\ E)$ is not accomodated, it follows that $S_1(A\ D)$ is not accomodated either.

Explanation. When assessing how well a tree with inner-node state assignments can explain a character state as due to inheritance and common descent, the correct measure is the number of independent accomodated pairwise similarities, not the total number of accomodated pairwise similarities. Consider a character in which 100 terminals have state 0 and another 100 state 1, and two trees on which the first 100 terminals occur in one region and the other 100 in two regions. Assume that in the first tree, the first region with state 1 has one terminal and the second 99; and that, in the second tree, both regions with state 1 have 50 terminals. The total number of pairwise similarities in this character state is $99 \times 100/2 = 4\,950$, of which at most 99 are independent. Summing over regions, in the first case a total of $0 + 4\,851 = 4\,851$ similarities are accomodated, in the second case only $1\,225 + 1\,225 = 2\,450$.

Yet in both cases, the same number of 98 independent pairwise similarities are required for a non-redundant description of the situation. Or, conversely, in both cases only a single independent pairwise similarity cannot be explained as a homology. This is in direct agreement with the observation that both cases can equally well explain the observations on account of the theory, which in this restricted case is possible historical identity of state 1 through inheritance and common descent on the given trees with the given sets of inner-node state assignments for the given character. The total number of pairwise homologies gives a different answer (the first tree is considered about twice as good: score 4 851 vs. 2 450) because that number also depends on the numbers of terminals that are present in each region of a tree in which the state is homologous. As these numbers do not feature in the theory on account of which the data are explained, the total number of accomodated similarities is not suited to measure agreement between theory and observation.

Weighting. An alternative way of viewing the difference between all and independent pairwise similarity statements is in terms of dynamic weighting of similarity statements (see De Laet and Smets 1998 for a similar discussion in the context of triplet and quartet methods). More particularly, if the weight that is assigned to an independent accomodated similarity statement in a given region is calculated dynamically as the total number of statements in that region divided by the number of independent statements in that region, then the total number of unweighed accomodated statements equals the number of weighted independent accomodated statements. This weighting scheme is highly unnatural and hard if not impossible to defend, which just reinforces the conclusion of the previous paragraph. But it also raises the general question of weighting.

I have been assuming equal weighting of similarity statements throughout, but the principle of parsimony as discussed here does in itself not prescribe that all parts of the data be equally weighted. Farris (1983, p. 11) discussed this issue at the level of differential weighting of entire characters and characterized his preference for equal weighting as a stance of ignorance: in the absence of any convincing reason for doing otherwise, all characters in a data set are treated as if they provide equally cogent evidence on phylogenetic relationship. The same reasoning applies at the level of the independent similarity statements that make up characters.

Algorithms such as Farris (1970; additive characters) or Sankoff and Rousseau (1975; step matrices) can be seen as methods that apply differential weighting within characters. Such differential weighting is defined in terms of transformations, not in terms of similarities: transformations between different pairs of character states can receive different weights. This may seem problematic for the current approach because the simple equivalence of minimizing homoplasy and maximizing homology, as discussed above, in general only holds when all transformations and all unit homologies are weighted equally. However, differential weighting as in Farris (1970) and Sankoff (1975) can also be characterized in terms of similarities that are hierarchically nested. A full discussion of this issue is beyond the scope of this review.

A methodological requirement. The unit of evidential value of a data set on a tree that arises from this discussion is an independent accomodated pairwise similarity statement. Likewise, independent pairwise similarity statements are the currency in which the empirical content of a data set is measured. This ultimately permits to interpret the preference for independent accomodated statements (versus all accomodated statements) as a methological requirement when maximizing the number of pairwise similarity statements that can be explained as homology: it enforces that each unit or quantum of empirical content of a data set is considered precisely once. Note that, in itself, this does not amount to equal weighting: whether or not all quanta of comparative empirical content should receive the same weight is an entirely different question.

Again, this methodological constraint is not met in quartet and triplet methods (De Laet and Smets 1998). Likewise, it is not met in methods that base the inference on a square matrix of pairwise distances among terminals, such as neighbor joining (Saitou and Nei 1987), for the simple reason that the required information to do so is not present in such matrices. To be sure, neighbor joining can in principle operate directly on character state data (Saitou and Nei 1987, p. 410), but such data sets are mostly reduced to square distance matrices first. In maximum likelihood methods such as Felsenstein (1981), the constraint is met. The difference with parsimony analysis is that in such methods the explanation of a similarity statement on a tree is based on integration over all possible inner-node state assignments, using stochastic models of character evolution and best-scenario branch lengths (see, e.g., Steel and Penny 2000 and Goloboff 2003 for a discussion). As seen above, when looking for best trees, parsimony analysis evades uncertainty as to the true historical status of a similarity statement that can be explained as a homology on a tree at an entirely different level, thus enabling it to remain largely agnostic about details of the processes of character evolution.

Maximizing the amount of homology

Given a data set of characters, one has to identify the tree or trees on which the highest number of independent compatible pairwise similarity statements can be explained as homology. This involves an optimization at two different levels. First, which is the highest number of such homology statements on a given tree? Second, given a procedure to solve the first problem, which is (are) the tree(s) on which this number is maximal?

The first problem can be tackled one character at a time because there are no logical interactions among the explanations of different characters (this is a fundamental assumption that is not met when inapplicables are present). Within a character, though, it cannot be tackled one state at a time because the explanation of any given state imposes methodological constraints on allowed explanations of the other states. As discussed above, such constraints are met when inner-node state assignments are taken into account, in addition to the observed states at the terminal nodes. Therefore, a crude solution for optimizing a character on a tree is to generate all possible sets of inner-node state assignments and to count the number of independent accomodated statements for each (three different possibilities, on the same tree, are illustrated in Figs. 6.2c–6.2e, with scores 5, 5, and 2). If the sets of inner-node state assignments are generated in a clever enough order, this can be improved using a branch-and-bound mechanism.

However, a much more efficient approach is possible, starting from the above observation that the number of independent compatible homologies and homoplasies for a character add up to a number that is tree-independent. As a result, a set of inner node state assignments that minimizes independent homoplasies also maximizes independent homologies. Next, the minimum number of independent homoplasies for a given character and a given optimal set of inner-node state assignments equals, up to a tree-independent constant, the number of regions as imposed by the inner-node state assignments, which in turn is one more than the minimum number of steps in the character. Therefore, algorithms that minimize the number of steps in such characters can be used to maximize homology. Examples are the algorithm of Farris (1970) for binary characters and additive multistate characters, or the algorithm of Fitch (1971; see also Hartigan 1973) for unordered characters.

The second problem is illustrated in the two trees of Figs 6.2b and 6.2f: even if the second tree can explain some characters better than the first tree (e.g. *c10*), the first tree is preferred because it provides a better explanation of the data as a whole. The problem of deciding whether a given tree is an optimal tree for the data at hand is NP-complete (Foulds and Graham 1982). Practically, this means that in general the only way to find the best tree(s) is the hard approach of examining all possible trees that exist for the given terminals, either explicitly or implicitly, by using a branch-and-bound approach (for which see Hendy and

Penny 1982). Unfortunately, the number of trees grows so extremely fast as the number of terminals grows (see, e.g., Felsenstein 1978b) that this approach is only feasible for relatively small numbers of terminals. Exactly how many terminals can be analysed in this way depends on the structure of the data set and on the computing power and time that is available, but as a rule of thumb it is somewhere between 15 and 25. So, when dealing with increasingly larger numbers of terminals, one is practically forced to restrict the tree search to increasingly smaller subsets of all possible trees, proportionwise. In doing so, heuristics such as branch swapping are used to make sure that no or little computing effort is wasted on trees that are manifestly not optimal (for a broader discussion and some developments beyond simple branch swapping see, e.g., Goloboff 1999; Moilanen 1999; Nixon 1999; Moilanen 2001).

Both levels of optimization are logically independent, even if they are in practice often tightly integrated in heuristic approaches (see, e.g., Goloboff 1996b for examples). One could do a tree search using any imaginable function that computes a number from a tree and a data set, and, heuristic uncertainty aside, the resulting trees would be optimal according to that function. Therefore, when comparing and evaluating different methods, it is sufficient to examine the meaning of the function used to evaluate any single tree.

6.2.5 Characters revisited

Summarizing this long introductory section, observation-based pairwise similarity statements are the fundamental statements of comparative research. When searching for trees on which the highest number of such similarities can be explained as homologies, two methodological requirements must be met: (1) the overall explanation of the data must be free of internal contradictions, which can be enforced by assigning, for each character, states to inner nodes of the tree; (2) the same piece of empirical content should not be used multiple times, which translates into counting only homologies that are logically independent.

From this point of view, a character that describes the distribution of a number of states in a number of terminals is just a convenient nonredundant summary of elementary putative homology decisions that are made, during character analysis, in all possible pairwise comparisons of some observable characteristic in those terminals (see De Laet and Smets 1998, pp. 378–380; the unhappy informal use of the term 'essence' does not invalidate their discussion). In each such pairwise comparison, the mere fact that the characteristic is being compared entails the hypothesis that at some level of generality it is historically the same. At a lower level, the different states of the character are hypotheses of alternative expressions of the characteristic, each of which is also hypothesized to be historically the same. As discussed above, all such hypotheses are to be seen through the lens of the Hennig–Farris auxiliary principle.

To clarify, consider some angiosperms and a character that codes a floral structure that comes in two forms, rounded (state 0) and square (1). The fact that these two forms are coded as states of the same character reflects the hypothesis that the structures, despite the observed difference in form, are homologous at a more general level. Mostly, such an hypothesis is based on a combination of criteria. As an example, when the development of floral buds in different terminals is compared, the meristem that gives rise to the structure could originate in almost identical topological relationships relative to other meristems. In addition, the adult structures, whether round or square, could share many anatomical and morphological similarities. As a whole, the character then reflects the higher-level prior hypothesis that the structure in all these terminals is identical through common descent and inheritance. Within the character, the difference in general form (round vs. square) is considered important enough to warrant recognition of two different states, reflecting the lower-level prior hypotheses that the roundness and the squareness of these structures can be explained as identity through common descent and inheritance as well.

The different roles of characters and character states

It has often been observed that there is a large discrepancy between the formalized nature of phylogenetic analysis once a data set has been constructed and the much more subjective decisions that are involved in character analysis, when it comes to deciding if observed features in two terminals should be coded as the same state of a character, two alternative states of a character, or part of different characters altogether (see e.g., de Pinna 1991, p. 380). Pleijel (1995) argued that this is especially relevant for the assumptions regarding homology of states within a character (are two such floral structures homologous, irrespective of their general form?). Contrary to hypotheses of homology within states (is the roundness of two round structures homologous, is the squareness of square structures homologous?), such higher-level hypotheses are never questioned during subsequent phylogenetic analysis (Pleijel 1995, p. 312). As an example, consider character *c9* of the data set of Fig. 6.1, and assume that state 0 codes the square and state 1 the round structure of the above character. On the most-parsimonious tree for these data (Fig. 6.2b), the squareness of the structure that is observed in terminals *out1* and *out2* is not homologous to the squareness of the same structure that is observed in terminals *E* and *F*, and the initial lower-level hypothesis has to be revised.

Similar posterior revisions of the higher-level hypothesis cannot be made because the homology of round versus square structures has been hard-coded in the analysis, precisely because they have been coded as states of the same character. To remove such hard-coded higher-level assumptions, Pleijel (1995) proposed to use absence/presence coding of character states, which is formally identical to non-additive binary coding, a technique that stems from phenetics (see, e.g., Sokal 1986). Whether it is feasible or desirable to exclude such assumptions from the analysis will be examined below.

But whatever the answer, the use of absence/presence coding as a means of doing so can lead to internal inconsistencies in the phylogenetic explanation of data, a result that is particularly relevant for this paper because Pleijel (1995) advanced absence/presence coding as a promising way to deal with inapplicables. Consider the data set of Fig. 6.3a and assume, without loss of generality, that none of the character states codes for absence. In the recoded version of Fig. 6.3b each column stands for one character state of a character of Fig. 6.3a, with 0 coding for absence of that state and 1 for presence. When analyzing Fig. 6.3a, the three trees of Fig. 6.3c are obtained (nine steps; loss of two independent pairwise similarities). With the recoded data, only one shortest tree is found, the middle tree of Fig. 6.3c; the two other trees are suboptimal by one step (18 vs. 17).

Pleijel (1995, p. 313) pointed out that, with absence/presence coding, hypotheses concerning

(a)

	c1	*c2*	*c3*	*c4*	*c5*	*c6*
out1	0	0	0	0	0	0
out2	0	0	0	0	0	0
A	1	1	0	0	1	1
B	1	1	0	1	1	1
C	1	0	1	1	1	2
D	1	0	1	1	0	2

(b)

out1	01	01	01	01	01	100
out2	01	01	01	01	01	100
A	10	10	01	01	10	010
B	10	10	01	10	10	010
C	10	01	10	10	10	001
D	10	01	10	10	01	001

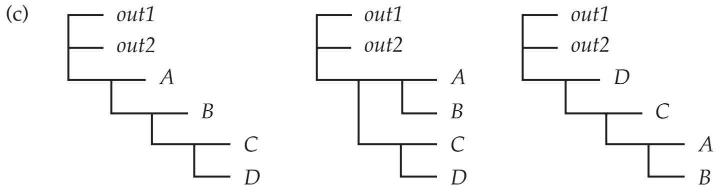

Figure 6.3 Absence/presence coding of character states aims to remove prior hypotheses of homology among states (Pleijel 1995) but can lead to internal inconsistencies. (a) A dataset with characters that reflect nested hypotheses of homology as determined during character analysis (characters unordered). (b) The characters of (a) with absence/presence recoding of character states. (c) The three most-parsimonious trees for (a). With the data coded as in (b) only the middle tree is considered optimal. The two other trees are rejected even if they explain the data equally well under acceptable hypotheses of homology that they imply.

transformation series between the analysed states will emerge as part of the results, but he remained somewhat vague about the logical and technical implications of this observation. As an example, take the three recoded states of the original character *c6*, each with a perfect fit on the single most-parsimonious tree for the recoded data. Because 0 stands for absence of the corresponding state, an inner node that is optimized as 0 can be hypothesized to have one of the two other states (other possibilities exist but are not relevant for the argument). Combining and summarizing all possible such optimizations of the three recoded states of *c6*, and using the outgroup hypothesis, three possible *implied transformation series* emerge from the tree: 1 ← 0 → 2, 0 → 1 → 2, and 0 → 2 → 1. Each of these has a perfect fit on the tree as well, and in each case only two steps are required to explain the state distribution. When doing the same excercise for the groups of states as defined by the other characters of Fig. 6.3a, all these other states can be explained by postulating a total of only seven steps (note that some of the implied transformation series incorporate non-homology of states as defined *a priori*; an example is character *c4*).

The middle tree of Fig. 6.3c is considered the best tree for the recoded states because it has the shortest length for the recoded data. But on the basis of possible transformation series that emerge as part of the analysis, one can construct a phylogenetic explanation of the data on that tree that requires fewer steps. So, whatever the length of an absence/presence recoded matrix on a tree means, it definitely does not measure how well that tree can explain the data phylogenetically under the assumption that character states can transform into one another, and maximization of phylogenetic explanatory power under that assumption cannot be the rationale for preferring trees that minimize this recoded length. Indeed, analyzing the two other trees in the same manner, they can also be explained by postulating only nine steps (which should not come as a surprise, as it was already clear from the analysis of the data set of Fig. 6.3a that the states could be grouped such that only nine steps are required on those trees). Yet they are rejected if the length of the recoded matrix is used as an optimality criterion.

(a)	*c1*	*c2*	*c3*	*c4*	*c5*	*c6*	(b) *c5′*	*c6′*
out	0	2	4	6	8	10	8	10
A	1	2	4	6	9	11	11	9
B	1	3	5	7	8	12	8	12
C	1	3	5	7	9	13	13	9

Figure 6.4 Absence/presence coding of character states, to remove prior hypotheses of homology among states, can lead to surprising optimal implied transformation series. (a) A dataset with six unordered characters as they return from character analysis; the groupings of character states in columns (characters) reflect nested hypotheses of putative homology; the most-parsimonious tree is (*out* (*A* (*B C*))), which is also the best tree when the data are recoded to remove prior assumptions of homologies among states. (b) Alternative grouping of the states of characters *c5* and *c6* that cannot be rejected on the basis of the optimized recoded states. For this grouping, the transformation series as implied by the optimized recoded characters provides a better explanation of the data than the original characters.

One step further, posterior groupings of states may exist that reduce the total number of steps below the number required by the groupings as they come out of character analysis. An example is presented in Fig. 6.4. As above, it can be assumed without loss of generality that none of the states in Fig. 6.4a codes for absence. When states 8–13 are grouped as in characters *c5* and *c6* of Fig. 6.4a, the transformation series that are implied by the optimizations of the recoded states on the best tree require a total of five steps on the best tree. But the alternative grouping as in Fig. 6.4b, implying 11 ← 8 → 13 and 10 → 9 → 12, can explain the observed distributions of states 8–13 at only four steps. This optimal implied grouping of states obviously contradicts the empirical evidence on the basis of which the original characters were proposed. But then it is the aim of this approach to remove such untestable assumptions (Pleijel 1995, p. 312), and posterior acceptance of groups of states as in characters *c5′* and *c6′* is just a logical consequence. More precisely, recognition of such transformation series follows from the notion that hypotheses concerning transformation series among the analysed states should emerge as part of the results and from the general requirements that the analysis should be logically capable of phylogenetic interpretation and internally consistent.

It does not require much imagination to see that in practice this could easily lead to situations where square floral structures of one angiosperm would *a posteriori* be considered homologous with, for example, a type of root system as present in another angiosperm, and the round floral structures of this other angiosperm to the root system of the first. Most systematists would not hesitate to reconsider homology within states on the basis of a well-supported most-parsimonious tree (the squareness of the floral structures in these terminals is not the same as the squareness of such structures in those other terminals after all, despite my prior assessment to the contrary), but in general such reinterpretations across characters are much more difficult to accept (darn, these flowers are actually not flowers but modified root systems!).

So, even if statements of homology among states are untestable in the sense of Pleijel (1995), they put bounds on the degree of reinterpretation of character states one is willing to accept in the light of incongruence in the data, and these bounds reflect empirical evidence as obtained during character analysis. Outright removal of such bounds, as would seem to be a logical consequence of using absence/presence coding as advocated by Pleijel (1995), therefore amounts to throwing away important relevant empirical data. As a work-around, one could limit implied transformation series to include only groupings of states that are compatible with the results of character analysis. But that actually amounts to giving up the premise that prior statements regarding homology among states should be removed from the analysis. And as discussed above, absence/presence coding then results in the same trees as obtained with regularly coded characters, at least if the aim of the analysis is to maximize explanatory power in a phylogenetic context.

Beyond single-column characters

On the other hand, it is not uncommon in character analysis to find multiple possible interpretations for features, which is not surprising given the role of background knowledge as discussed earlier. As an example, depending on the view one takes, the vegetative region in some species of the angiosperm genus *Utricularia* (bladderworts) can be interpreted morphologically as a shoot-like leaf, a branched stem system without leaves, or a shoot with stems and leaves (Rutishauser and Sattler 1989; a fourth, more complex, interpretation is also provided). Similar problems abound when dealing with fossils or when making comparisons across very divergent groups. In both cases one often has to deal with structures that cannot be easily homologized across the terminals being compared, which in turn often results in competing and conflicting prior interpretations. In studies of sequence data, this problem can come in the form of different prior hypotheses about orthology and paralogy of sequences (Fitch 1970) or in different alignments for the same set of putative orthologs (several examples of the latter case are discussed in the second section).

In each such case, when characters are coded according to just one of the competing interpretations, chances are that the chosen view will be favored by the resulting trees simply because the data have been exclusively interpreted as such to begin with. As observed by Endress (1994, p. 401–402), circular reasoning when dealing with such ambiguously interpretable features can be overcome by repeatedly testing all different possibilities. Only this approach amounts to a sincere attempt at falsification. Unfortunately, in formal analyses and with current algorithms this is not easy to achieve because the technical framework of independent single-column characters does not lend itself to simultaneous analysis of such alternative interpretations of the data in a logically consistent and correct way.

A hard work-around would be to manually construct and analyse as many data sets as there are different combinations of different interpretations in different characters, which may be practically feasible when the number of such combinations is not too large. The best phylogenetic hypotheses would then be the shortest trees across all those data sets, and optimal homologizations and details of transformation series would emerge from those trees as part of the analysis. The difference with absence/presence coding of states is that, as above, the level of reinterpretation of states that one is willing to accept in the light of incongruence is still bounded by the results of character analysis. The difference with an analysis of just one set of classic single-column characters is of a purely technical nature: these are cases in which the *a priori* acceptable hypotheses of homology among states cannot be expressed as a simple series of independent single-column characters. But the purpose remains maximization of the number of independent pairwise similarities that can be interpreted as identical through common descent and inheritance. From this point of view, the next section can be seen as an attempt to develop a formal and logically consistent method to deal with the problem of multiple *a priori* acceptable hypotheses of homology among states in the case of putative homology statements within putative orthologous sequences.

6.3 Parsimony analysis of sequence data

When dealing with sequence data, it is not unusual to find that putative homologous sequences have different lengths in different terminals. Such length differences are explained as the result of indel events, insertion and/or deletions that occurred in the course of evolutionary history. As a consequence of indel events, two sequences that are homologous as a whole will nevertheless contain subsequences that are not homologous: with a deletion, the resulting sequence misses a part of the original sequence; with an insertion the resulting sequence has a subsequence that was not present before. In both cases, characters that describe the subsequences that are involved will be inapplicable in the other sequence.

For the purpose of phylogenetic analysis, it is common practice to establish the positions and sizes of indels by creating a multiple alignment prior to tree evaluation and tree search, thus turning the putative homologous sequences into a sequence of single-column positional characters that subsequently can be treated as a regular data set (see Fig. 6.5a for an example). Each such positional character describes the state distribution of the base that is found at that position of the alignment, with gaps (coded as dashes in this chapter) indicating inapplicability. As discussed by Maddison (1993, p. 578), this makes sequence data susceptible to the general problems that come with inapplicables.

However, the approach of generating multiple alignments prior to tree evaluation and tree search is fundamentally insufficient as a general method for analysis of sequence data, as will be discussed below. As a consequence, the question of inapplicables in sequence data cannot be discussed in

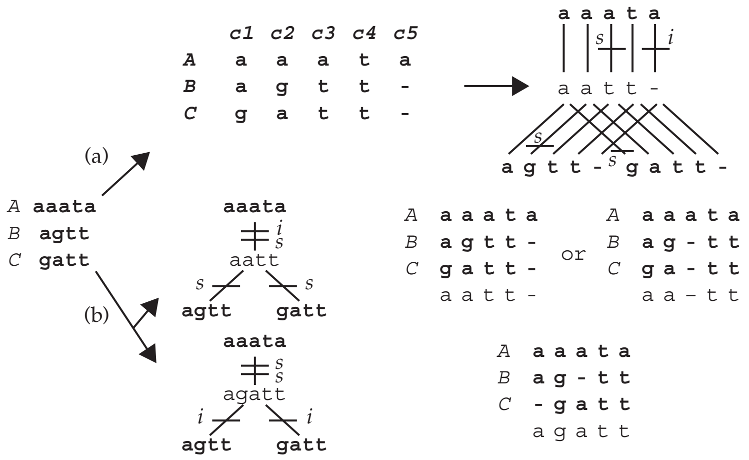

Figure 6.5 Three putative homologous sequences and two different approaches to evaluating them on the single unrooted tree for three terminals. (a) First a multiple alignment is constructed to establish base-level positional correspondences (dashes indicate gaps); the resulting positional characters are optimized using the algorithm of Fitch (1971), resulting in three substitutions (*s*) and one indel (*i*). (b) The unaligned sequences are optimized directly on the tree using the algorithm of Sankoff (1975); in this example, two optimal reconstructions of the sequence at the inner node exist, each at four steps; in each case, the optimal length imposes one or more optimal sets of positional correspondences.

general at that level. It is argued that a general method by necessity requires that unaligned sequences be directly optimized on trees, using algorithms such as Sankoff (1975) or Altschul (1989, pp. 307–308). Such algorithms treat the unaligned putative homologous sequences as one single complex character, to which I shall refer as a *sequence character*. It is widely believed that the various parameters that these algorithms employ to set up a cost regime, such as base substitution and gap costs, can only be specified or interpreted with reference to detailed models of the evolutionary processes that generated the data. However, the cost regime can also be set according to the principle of parsimony as discussed above, leading to a maximization of the amount of independent sequence similarity that can be interpreted as due to inheritance and common descent (De Laet 2004).

Throughout this section I use DNA sequences, but the discussion is general and applies to any kind of data that can be conceptualized to be hierarchically related through substitutions and indels, including, for example, serial homologs in morphology or different versions of manuscripts in stemmatology. Examples are constructed such that optimalities can be verified by hand.

6.3.1 Some background

Some additional notes on terminology are appropriate first. Gap and gap cost terminology can be confusing because the same terms are sometimes used for different things and the other way around. As an example, in a sequence like *a t t - - - t t a c* the term gap is sometimes used for each of the three consecutive missing positions in the middle (three gaps), or alternatively for the whole stretch of three missing positions (one gap). In this paper, a *gap* always refers to a maximum stretch of missing positions, not to smaller composing parts. The *length* of a gap is the number of positions over which it extends. The smallest composing part of a gap is referred to as a *unit gap*. The character that is used to indicate a unit gap, a dash in this chapter, is sometimes called the *gap character*, a term that has also been used for characters in data sets that describe the distribution of a putative indel events (e.g. Simmons and Ochoterena 2000).

All gap costs in this paper are of the form $a + (n - 1) * \mathrm{b}$, in which n is the *length* of the gap, a the *(gap) opening cost*, and b the *(gap) extension cost*. If gap opening cost and gap extension cost are equal, the term unit gap cost refers to either, and the cost for a gap of length n is n times the unit gap cost. Such a cost regime can be expressed as a 5×5 step matrix (see Sankoff and Rousseau 1975) in which the unit gap is included as a fifth state, in addition to a, c, g, and t.

The *minimal mutation algorithm* of Sankoff (1975) is illustrated in the example of Fig. 6.5b. It reconstructs *inner node sequences* and *positional correspondences* among observed sequences such that the total number of mutations is minimized under the assumption that a gap of length n constitutes n mutation events. This corresponds to a cost regime in which all base substitution costs, the gap opening cost, and the gap extension cost are equal. Sankoff and Cedergren (1983) generalized the approach to a step matrix with arbitrary metric distances, still treating a gap of length n as n events. A further extension to include gap costs of the form $a' + n * b$, in which n is the length of the gap, $a' + b$ the gap opening cost, and b the gap extension cost, was examined by Altschul (1989, pp. 307–308). With such gap costs, the first unit gap of a gap incurs a cost $(a' + b)$, each next unit a cost of b.

Sankoff (1975) used the concept of optimal *frame sequences* to specify reconstructed sequences and positional correspondences that lead to minimal costs. Sankoff and Cedergren (1983) framed their discussion in terms of the slightly less general concept of *tree alignments*. A tree alignment always refers to a particular tree with the given sequences at the tips and hypothetical or reconstructed sequences at the inner nodes. It consists of (1) that tree; (2) a matrix in which both observed and reconstructed sequences are aligned; and (3) correspondences between nodes of the tree and rows of the matrix. It is conveniently represented as a tree in which the nodes are labeled with the rows of the matrix, as, for example, in Fig. 6.10 (see below). In this way it is easy to see that, in a tree alignment, each branch of the tree defines a pairwise alignment between the sequences at the two nodes that the branch connects. The *cost of the tree* *alignment* is then defined as the sum of the costs of these pairwise alignments along all branches of the tree, always with reference to the cost regime in use. A 'classic' multiple alignment of the terminal sequences is obtained by deleting the rows with inner-node sequences from the matrix of a tree alignment. Multiple alignments that are obtained in this way have been called *implied alignments* (e.g. Schwikowski and Vingron 1997; Wheeler 2003a). Some examples of optimal implied alignments can be found in Fig. 6.5b.

With cost regimes that make no difference between gap opening cost and gap extension cost, the cost at any position in a pairwise alignment of a tree alignment is independent from the costs at its other positions. By extension, this also applies to the costs of complete colums of a tree alignment. As a result, each such column can be interpreted as a single-column character with a set of inner-node state assignments. In this way the algorithm of Sankoff (1975; all substitution costs and unit gap cost equal) can be seen as a generalization of the minimum mutation algorithm of Fitch (1971). Indeed, under the conditions of Sankoff (1975), each column of an optimal tree alignment specifies a character and set of inner-node state assignments that are also optimal under the conditions of Fitch (1971). The generalization lies in the fact that different optimal tree alignments for the same data on the same tree can imply different sets of Fitch characters (see Fig. 6.5b for examples). The algorithm of Sankoff and Cedergren (1983; tree alignments with step matrices) is a similar generalization of the algorithm of Sankoff and Rousseau (1975), which, in turn, generalized Fitch (1971) to accomodate differential weighting within characters. Under the conditions of Altschul (1989; different gap opening and gap extension costs), the costs of the different columns of a tree alignment are no longer independent. As a result, such tree alignments cannot be understood in terms of independent single-column positional characters.

As was the case with inner-node state assignments for simple single-column characters (compare, e.g., Figs. 6.2c and 6.2e), tree alignments on a given tree can be optimal or suboptimal. Sankoff and Cedergren (1983) called the cost of an *optimal* tree alignment for a set of observed sequences on a given tree the *tree distance* of those sequences on that tree. Their and similar algorithms (Sankoff 1975; Altschul 1989) can be used to calculate such tree distances and the reconstructions that come with them. In terms of the current approach, the tree distance as defined by Sankoff and Cedergren (1983) is the *length* of the sequence character on that tree. As such, the algorithms of, for example, Fitch (1971) and Sankoff (1975) are comparable in the sense that they both calculate the cost of an optimal reconstruction of a character on a tree. As will be discussed below, they are vastly different when it comes to computational complexity. For tree alignments, the second level of optimization—the problem of finding, among all possible trees, trees of minimal length or tree distance—is often called *generalized tree alignment* (e.g. Jiang and Lawler 1994; Vingron 1999) but other terms are used as well; Hein (1989a), for example, refers to it as the *general parsimony problem*.

6.3.2 Putative homologous sequences: a sequence of characters or a sequence character?

It has been argued that all substitution costs and the unit gap cost should be set equal in Sankoff (1975) style analyses of sequence data (Frost *et al*. 2001), a position that will be examined more closely later. However, first it is argued, in this subsection, that a general method of sequence alignment must by necessity move beyond prior multiple alignments (contra Simmons and Ochoterena 2000; Simmons 2004). The argumentation does not depend on the particular settings of the cost regime, but for clarity I tentatively accept the position of Frost *et al*. (2001) and contrast (equally weighted) Fitch (1971) analysis of prior alignments with Sankoff (1975) analysis of unaligned sequences.

When optimizing a sequence character on a tree, base-level correspondences among the observed sequences are not determined and fixed *a priori* but calculated as part of the optimization process, as already illustrated for three terminals in Fig. 6.5. The full implication of this can be seen when analyzing more than three sequences, such that alternative trees exist and have to be examined. Consider the data set of Fig. 6.6a. For four taxa

```
(a) A  gc     (b) A  gc     (c) A  gc     (d) A  gc-
    B  cg         B  cg         B  cg         B  cg-
    C  c          C  -c         C  c-         C  --c
    D  gg         D  gg         D  gg         D  gg-
```

Figure 6.6 A simple dataset (a) and three different multiple alignments (b, c, d).

A–*D*, three unrooted trees exist: (*A B*)(*C D*), (*A C*)(*B D*), and (*A D*)(*B C*). Using Sankoff (1975), the latter two are both diagnosed at cost 3 (each time two substitutions and one indel) while (*A B*)(*C D*) comes at cost 4 (three substitutions and one indel). Looking at the two optimal trees, (*A C*)(*B D*) comes with the implied alignment of Fig. 6.6b, (*A D*)(*B C*) with the different implied alignment of Fig. 6.6c. So it is not just that base correspondences are not fixed prior to analysis, *a posteriori* they can be different in different optimal trees.

A simple case of symmetry

The data set of Fig. 6.6a has a peculiar symmetry: when the labels of *A* and *B* are switched and the directions of all sequences reversed, the original data set is recovered. As such it provides a perfect example where mutually exclusive sets of putative homology statements cannot be distinguished at the level of character analysis. The higher-level hypothesis in this data set is that the sequences are orthologs. Within the orthologs, however, the symmetry makes it logically impossible to decide *a priori* if the single *c* of terminal C is to be considered homologous to the *c* in the second position of *A* or to the *c* in the first position of *B*. Conceptually, this is like the situation in bladderworts, discussed above, where it cannot be determined *a priori* if the vegetative system should be considered a shoot-like leaf or a leaf-like shoot system (even if the situation with bladderworts is more complex because there are still other homologizations that are considered acceptable on *a priori* grounds).

Turning to trees, the symmetry has, as a consequence, that these data cannot possibly distinguish between (*A C*)(*B D*) and (*B C*)(*A D*), two unrooted trees in which the labels of *A* and *B* have been exchanged. This conclusion follows directly and solely from the internal structure of the data set. As such it can be used to establish the following strong test for candidate phylogenetic methods: (*A C*)(*B D*) and (*B C*)(*A D*) should get the same score. Any method that does not meet this test is in serious trouble.

As discussed, Sankoff (1975) optimization diagnoses (*A C*)(*B D*) and (*B C*)(*A D*) at the same cost and thus meets the test. Turning to prior alignments, the first question is which prior alignments to consider. With data as simple as this it is easily established that alignments in Figs 6.6b and 6.6c are the only valid candidates. All other alternatives, such as, for example, Fig. 6.6d would need some special argumentation as to why, in this case, the *c* that is observed in terminal *C* should not *a priori* be considered homologous to the *c* that is observed in *A* or to the *c* that is observed in *B*. Given that it is accepted, *a priori*, that the sequences as a whole are homologous (they are putative orthologs), this seems hard to do. A Fitch (1971) analysis of alignment 6b yields tree (*A C*)(*B D*) at cost 3, with (*B C*)(*A D*) one step more costly; alignment 6c yields (*B C*)(*A D*), also at cost 3, and with (*A C*)(*B D*) one step more costly (in both cases, (*A B*)(*C D*) has a cost of 4). So, when looking at just one alignment, the two trees get a different score and the method fails the above test. As a result, depending on the prior alignment that is used, positive support is found for either (*B C*)(*A D*) or (*A C*)(*B D*), whereas in fact relationships are ambiguous.

Similar symmetry observations can be made with respect to alignments 6b and 6c: they can be turned into one another by exchanging the labels of *A* and *B* and reversing the direction of each sequence. Therefore, if either is considered optimal according to some criterion, the other should be as well. So a way out of the problem of finding spurious relationships with single prior alignments suggests itself: rather than to construct and analyse just one prior alignment, identify and analyse all different prior multiple alignments that are considered optimal, and accept only groups that are common to all. This may sound trivial but it raises the non-trivial question of how to calculate the relevant prior optimal multiple alignments. For this particular example, that question comes down to finding a criterion that gives an optimal score to alignments on Figs 6.6b and 6.6c and a worse score to all other alignments.

Optimal alignments of two sequences can be calculated using dynamic programming algorithms as pioneered, in biology, by Needleman and Wunsch (1970) and Sellers (1974). A description of the basic algorithm and some historic notes can be found in Kruskal (1983); extensions are reviewed in, for example, Gusfield (1997). For the current purpose, approaches that generalize such algorithms to more than two sequences can be grouped according to whether or not they use the tree-alignment approach.

In optimal tree alignments, the kind of data symmetry in Fig. 6.6a is reflected directly in symmetry of calculations when comparing trees (*A C*)(*B D*) and (*B C*)(*A D*). So it was not just coincidence that the above Sankoff (1975) optimization of the data of Fig. 6.6a gave identical scores for those trees, with implied alignments that display among themselves the same symmetry as the data. Theoretically then, one could use a tree-alignment analysis to generate implied alignments that are next used as prior alignments. There would be no need to analyze the implied alignments, though, because their best trees would already have been identified in the preliminary tree alignment analysis. In fact, while the approach provides a solution to the problem discussed here, it actually comes down to giving up the notion that sequences should be aligned prior to tree evaluation and tree search.

Among the multiple alignments methods that do not use tree alignments, *SP alignments* or *sums-of-pairs alignments* (Murata *et al*. 1985; Carillo and Lipman 1988) and especially *progressive alignment* methods (e.g. Feng and Doolittle 1987; Thompson *et al*. 1994; Notredame *et al*. 2000) are probably most widely used. First consider SP alignments. An SP alignment of a set of sequences is an alignment for which the sum of pairwise alignment scores between all possible pairs of sequences is minimal. Setting all substitution costs and the unit gap cost to 1, it is easily verified that the alignments of Figs 6.6b and 6.6c have identical SP scores of 9, leaving the SP criterion as a potential solution to the problem.

```
(a) A  ct    (b) A  ct    (c) A  ct
    B   c        B  -c        B  c-
    C  tc        C  tc        C  tc
    D  tc        D  tc        D  tc
    E  tt        E  tt        E  tt
```

Figure 6.7 A simple data set (a) and two different multiple alignments (b, c). According to the SP criterion, alignment (b) is better than alignment (c) (SP scores 13 and 14).

Another case of symmetry

However, consider the data of Fig. 6.7a. Reading each sequence in reverse, nothing changes for *B* and *E*, but the sequence of *A* is turned into the sequence of *C* and *D*, and the sequences of *C* and *D* are turned into the sequence of *A*. Therefore, the structure of the data set is such that these data cannot distinguish between trees that differ only in the positions of *A* vs. (*C D*), as, for example, the pair (*B* (*C D*) (*A E*)) and (*A B* (*E* (*C D*))). Using Sankoff (1975), these trees both have a cost of 3, which is the optimal cost over all trees as well. Tree (*B* (*C D*) (*A E*)), or any other tree that has an *AE*–*BCD* partition, comes with optimal implied alignment 7b; tree (*A B* (*E* (*C D*))), or any other tree that has an *AB*–*CDE* and an *ABE*–*CD* partition, comes with alignment 7c. As above, these implied alignments have among themselves the same symmetry as the unaligned data. So Sankoff (1975) optimization does not tell these trees apart, and correctly so.

This is necessarily so as long as the ancestor of *C* and *D* has a reconstructed sequence that is identical to and perfectly aligned with the sequences of *C* and *D* in optimal tree alignments. If this is the case, the data symmetry is directly reflected in the Sankoff (1975) calculations that are performed on the two trees that are involved, and an identical cost on both trees follows. The assumption about the reconstructed sequence for the ancestor of *C* and *D* is easily proved by showing that its negation leads to a contradiction. Assume that an optimal tree alignment exists in which the ancestor of *C* and *D* has a sequence that is different or differently aligned. In that case, the tree alignment can be improved—contradicting the premise—by changing that ancestor and its alignment as indicated above. That this is an improvement can be seen as follows: for any position in the ancestor of *C* and *D* with an entry (base or unit gap) that is different from the base at the corresponding position in *C* and *D*, changing that entry into the corresponding entry of *C* and *D* will improve the cost by two mutations; at the same time, that change can incur at most one additional mutation, between the ancestor of *C* and *D* and the third node to which this ancestor is connected. So, in conclusion, optimal tree alignments are not tricked by data symmetries such as in Fig. 6.7.

This does not hold for SP alignments: alignment 7b has a better SP score than alignment 7c (13 vs. 14; the score for 7b is optimal), proving the case by counter example. As a result, if the SP criterion were used to construct and select prior alignments, alignment 7b would be selected and trees with *AB–CDE* and *ABE–CD* partitions considered suboptimal in the subsequent phylogenetic analysis. To salvage the approach, one could consider to examine suboptimal SP alignments, like 7c, up to the degree that all prior alignments have been accepted that are involved in symmetries such as in Figs. 6.6a and 6.7a. But this would not work, for two reasons. First, there is no general way to tell how far one has to descend into suboptimality before all relevant alignments have been taken into account. Second, many additional and unwanted alignments might pass as well. So accepting suboptimal SP alignments cannot be a general solution to this problem of data symmetry.

Similar problems can arise with progressive alignments using *guide trees* (e.g. Thompson *et al*. 1994; see also Feng and Doolittle 1987). Such trees are usually constructed on the basis of a square overall distance matrix that is derived from pairwise alignment scores. Multiple alignment then proceeds by traversing this tree from terminals to the root. At each node that is visited, a partial multiple alignment is created that includes and combines the partial alignments that are found at the daughter nodes (terminal nodes are initially assigned a trivial partial alignment that includes just the observed sequence of that node). In this way, all sequences are included in the alignment after the root node has been visited. At any node, the alignment of partial alignments mostly proceeds by using some modification of the SP criterion, considering only those pairwise alignments across the node being considered. Moreover, this criterion is mostly applied only locally: gaps that have been inserted before will never be removed. In general, this group of methods cannot guarantee that symmetries as discussed here are properly taken into account.

A case of local symmetry

Based on the premise that multiple alignments should be constructed prior to tree search on the basis of a similarity criterion, Simmons (2004, p. 876; see also Ochoterena 2004) recently proposed the following tree-independent procedure for constructing optimal prior alignments. In a first step, construct one or more multiple alignments using, for example, programs that try to maximize (an unspecified measure of) similarity, or information from secondary structure. Next, evaluate these alignments using the number of 'differences' that are implied, and try to lower that score by adjusting those alignments. Such adjustments can be done manually or, ideally, using optimization programs. The rationale is to further increase the amount of similarity that is present in the alignment. The best alignments that are obtained are then subjected to parsimony analysis.

In the above, the number of differences is best explained by first looking at a regular data set such as in Fig. 6.1. For each character in the data set, the observed variation *m* (Farris 1989a, p. 417) is one less than the number of states in the character, and that number is the minimum of steps that the character can have on any tree. The observed variation for the data set as a whole, *M*, is the sum of the observed variation in all its characters, and can be interpreted as the number of steps that the best tree for the data set would have if all characters were congruent. If indel events would not occur, the number of differences in the sense of Simmons (2004) would be equal to *M*. But indel events do occur and complicate matters because single indel events can affect multiple columns of an alignment. However, as will be clear below, further details of the calculations that are involved in such cases (see, for example, Simmons and Ochoterena 2000) are not required for the current argument. Simmons (2004) observed that minimization of differences in this sense can lead to trivial alignments that require only as many indels as there are sequences in the data set, irrespective of the tree being considered (see Fig. 6.13c, below, for an example). To circumvent that problem,

```
(a) out  ttttttttttggggtttt tcca   (b) tcca   (c) tcca   (d)
    A    aattttttttggggtttt c          -c--       --c-
    B    aaaattttttggggtttt c          -c--       --c-
    C    aaaaaattttggggtttt c          -c--       --c-
    D    aaaaaaaattggggtttt c          -c--       --c-      I c
    E    aaaaaaaaaaggggaaaa cg         -cg-       -cg-      E cg
    F    aaaaaaaaaacccctttt gc         -gc-       -gc-      F gc
    G    aaaaaaaaaaccccaaaa aca        -aca       aca-      G aca
    H    aaaaaaaaaaccggaatt gg         -gg-       -gg-      H gg
```

Figure 6.8 An example of localized data symmetry. (a) A data set consisting of two sets of putative homologous sequences. (b, c) Two multiple alignments for the second set. (d) Reduced data set that exhibits the same kind of symmetry as discussed for Fig. 6.6.

Simmons (2004, p. 876) suggested not to add positions to alignments as obtained in the first step during possible adjustments in the second step.

This optimality criterion assigns the same scores to the symmetric alignments of Figs. 6.6 and 6.7, and in each case all other alignments have a worse score. Therefore this approach could correctly identify the relevant prior alignments for these problematic data sets. However, consider the data set of Fig. 6.8a, a case where two different sets of putative homologous sequences are analysed simultaneously (the example uses two sets of sequences for reasons of clarity only; similar examples can be constructed that use only one set of putative homologs). The structure of the first set of sequences jumps out so clearly that it is easily seen that the best trees for that part of the data are (*out* (*A* (*B* (*C* (*D* (*E* (*H* (*F G*)))))))) and (*out* (*A* (*B* (*C* (*D* (*F* (*H* (*E G*)))))))). Moreover, it is easily established that the first set of sequences is so strongly structured that the problem of finding the best trees for the data set as a whole reduces to evaluating the second set of sequences on those two trees.

In both trees, consider the ancestor of terminals *D*–*H* and this second set of sequences. In each case, that node will be optimized as *c* for the alignments of Figs. 6.8b and 6.8c, or indeed for any alignment in which the *c*'s of terminals *A*–*D* are aligned (it is easily seen that such must be the case for optimal explanations). Next consider the data set of Fig. 6.8d, where terminals *out*, *A*, *B*, *C*, and *D* have been replaced by a single hypothetical terminal *I* that is assigned that reconstructed sequence *c*. This reduced data set exhibits the same kind of data symmetry as discussed above: change the labels of *E* and *F*, reverse the direction in which the sequences are read, and the original data set is recovered. Considering all this, the second set of sequences of Fig. 6.8a cannot be used to distinguish between the two candidate trees, as these only differ in their relative positions of *E* and *F*. Therefore, any method that assigns different scores to these trees for these data is in serious trouble.

The algorithm of Sankoff (1975) properly takes into account data symmetries such as in Fig. 6.8d. It also treats the whole data set of Fig. 6.8a correctly, which can be shown, as above, by observing that optimal tree alignments on optimal trees have to reconstruct the ancestral sequence for terminals *D*–*H* as *c*, and such that this *c* is aligned with the *c*'s of terminals *A*–*D*. The score for the complete data set of Fig. 6.8a on both trees is 30, and this is also the optimal score. Two corresponding implied alignments are shown in Figs. 6.8b and 6.8c. As above, these display the same symmetry as the raw data (other optimal tree alignments exist, but that does not affect the argumentation).

Evaluating these implied alignments using the criterion of Simmons (2004) cannot be done by simply summing over isolated columns because some gaps affect more than one column, and more elaborate calculations are required. However, these are not really required in this case because reversing the sequences in both alignments establishes mutual symmetry of gap positions for such calculations. So, whatever the contribution of the gaps in the first alignment, it will be the same in the second and their unit gaps can therefore be treated as missing entries for the purpose of assessing the relative scores of the alignments. This results in relative score three for Fig. 6.8b but four for Fig. 6.8c, and the procedure of Simmons (2004) therefore would lead to prior rejection of the alignment of Fig. 6.8c. The net result is that this procedure leads to rejection of a tree that the data cannot distinguish from a tree that it accepts.

Comparing the alignments of Figs 6.8b and 6.8c, the preference of the optimality criterion of Simmons (2004) for the first one boils down to the fact that it puts the last *a* of terminal *G* in the same column as the last *a* of the outgroup. But on the best tree for this alignment, the *a* that *G* and the outgroup share cannot be explained as identical by common descent and inheritance. Consider the consequences of this observation in the light of the overall analysis, where tree (*out* (A (*B* (*C* (*D* (*E* (*H* (*F G*)))))))) is accepted but (*out* (A (*B* (*C* (*D* (*F* (*H* (*E* G)))))))) rejected. Given the local symmetry in the second sequence character, both trees explain the data equally well, albeit with different posterior homologizations of positions and base identities. But they are different in their amounts of homoplasy: overall, the first tree has a homoplastic pairwise base similarity (the last *a* of terminal *G* and the outgroup) that the second tree lacks. Moreover, the preference for the first tree when using the procedure of Simmons (2004) is based solely on this difference: of the two trees with equal amount of similarity that can be explained as homology, it selects the tree that has the higher amount of homoplasious similarity. In more complex cases, this effect can ultimately lead to rejection of trees with higher amounts of homologous similarity in favour of trees with lower amounts of homologous similarity. The same problem can also occur with the related tree-independent optimality criteria for multiple alignments that have recently been discussed by Carpenter (2003, pp. 6–7) and Nixon and Little (2004).

General conclusions

None of this is accidental. Data symmetries such as in Figs 6.6a, 6.7a, and 6.8a have a consequence that no distinction can be made between particular trees or groups of trees. As a result, methods of analysis that do not directly take into account the structure of trees (e.g. SP alignment or the procedure of Simmons 2004), or do so in a way that violates the symmetry (e.g. progressive alignment, or even just the use of suboptimal tree alignments), will not in general be able to deal with such situations. This leaves, by definition, optimal tree alignment methods. As a corollary, unless one is willing to defend methods that in some cases can give different scores to trees that cannot be distinguished by the data at hand, alignment and tree search cannot be properly separated in phylogenetic analysis of sequence data. Note that this conclusion is argued and reached in logical space. Whether or not it results in a practically feasible method will be discussed below.

The examples of Figs 6.6a, 6.7a, and 6.8a are unusual in that some terminals have sequences that are the exact reverse of other sequences, a situation that will hardly if ever arise in real data sets. But such perfect crab canons are not necessary for the phenomenon to occur. Sequences such as those can be embedded as short motifs in longer sequences that as a whole are not identical when read in reverse, and similar distortions could result. For simple examples as above, one could argue that the problem can easily be spotted and solved by carefully inspecting the data and the alignments by eye, but this approach would no longer work in such more complex cases.

In addition, the motifs that are involved do not have to be identical when read in reverse, only their alignment scores with the other sequences must remain unchanged. Lastly, even when the symmetry in the motifs is not perfect, by deviations in motif sequence and/or substitution costs that are involved, systematic distortions, though less well defined, would still arise. So situations where short subsequences can have alternative optimal alignments, with different local costs on different trees, may well be relatively common in empirical data. Moreover, when such data sets are aligned progressively according to a guide tree (using, for example, CLUSTAL; Thompson *et al*. 1994), such ambiguities that include groups of the guide tree may systematically be resolved in favor of the guide tree.

Summarizing, alignment and tree evaluation cannot be *properly* separated in phylogenetic analyses of sequence data. As a consequence, the view that a set of sequences that are deemed putative homologues should be turned into a sequence of positional characters prior to tree search and evaluation is erroneous or at best incomplete. Instead, such sequences constitute a single complex character, a sequence character, that can be optimized on trees using optimal tree alignment algorithms such as that of Sankoff (1975). These conclusions follow from very general considerations of data symmetry and do not depend on details of the cost regime that is used.

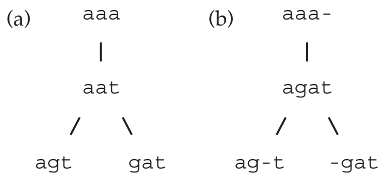

Figure 6.9 Two different tree alignments of the putative homologues *aaa*, *agt*, and *gat* on the single tree for three sequences.
(a) This reconstruction requires three steps (three substitutions, no indels) and retains three independent pairwise base similarities among observed sequences. (b) At four steps (one substitution, three indels) this reconstruction requires one more transformation, even if it retains one more independent pairwise similarity among observed sequences.

6.3.3 Quantifying and maximizing homology in sequence characters

Frost *et al*. (2001, pp. 354–355; they use the term 'indel' for a unit gap as used here) discussed the method of direct optimization (Wheeler 1996), and argued for setting all substitution costs and the unit gap cost equal because this amounts to equal weighting of all hypothesized transformations, which in turn 'renders the highest degree of descriptive efficiency and maximizes the explanatory power of all lines of evidence (i.e. characters).' Direct optimization has been proposed and is still often discussed as a sequence optimization method that is qualitatively different from optimal tree alignment methods, but the method is best seen as a heuristic approximation for optimal tree alignments (De Laet and Wheeler 2003; see also below), and the claimed novelty of the approach rests on a lack of familiarity with or misunderstanding or misrepresentation of the work of Sankoff (1975) and Sankoff and Cedergren (1983) (see, e.g., Wheeler 1996, 1998; Giribet and Wheeler 1999; Phillips *et al*. 2000; Wheeler 2001b, 2002, 2003a). Therefore, the argumentation of Frost *et al*. (2001) amounts to a preference for the minimum mutation algorithm of Sankoff (1975).

Consider the sequence character *aaa*, *gat*, and *agt* and two alternative tree alignments on the single tree for three terminals as presented in Fig. 6.9. With the above cost regime, tree alignment 9a is better than 9b (three steps versus four). On the other hand, when looking at independent accommodated pairwise similarities, as a measure of the amount of similarity that can be explained as homology, 9b performs better than 9a: it accomodates one more independent pairwise base match. This should not come as a surprise. For pairwise alignments, Smith *et al*. (1981; their equation 4b with $w_k = 0$) showed that maximization of base-to-base matches is equivalent to minimization of cost when all base substitution costs are set at twice the unit gap cost, a different regime than advocated by Frost *et al*. (2001). This result of Smith *et al*. (1981) cannot directly be extended to comparisons of more than two sequences, but a generalization to tree alignments (see below) still yields a cost regime that is different from the one favored by Frost *et al*. (2001). With more than three sequences, this difference can lead to a preference for different trees.

On a general level, this example merely reflects the well-known fact that the choice of substitution, gap opening, and gap extension costs affects the result of alignment and tree-building procedures. When examining the logical basis of sequence analysis, however, the paradoxical situation arises that the objectives of maximizing explanatory power and maximizing independent homologous similarity seem to be at odds. As discussed below, this contradiction is only apparent because the premises at either side of the comparison are faulty: setting all costs equal does not maximize explanatory power, and independent base-to-base homologous similarity is not all there is to sequence homology.

Subsequence homology and compositional homology
The latter is easily seen when considering a data set, such as in Fig. 6.10, where sequences differ only in length. The two tree alignments that are shown do not differ in the number of independent

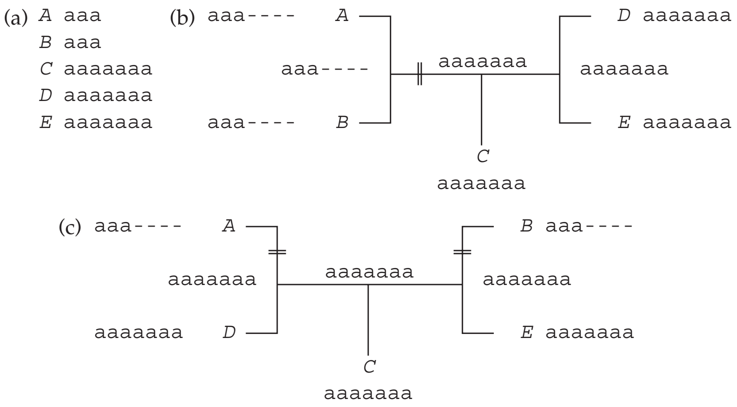

Figure 6.10 A data set in which the sequences only differ in their lengths (a) and two trees with optimal inner-node reconstructions and positional correspondences under the assumption that insertion/deletion of a stretch of contiguous bases is counted as one transformation (b, c). Double bars indicate indel events. Note that on each tree alternative sets of optimal positional correspondences exist.

base-to-base matches among observed sequences that they accommodate: in both cases there are 20 independent base-to-base comparisons, and all these are matches. Yet, the first tree alignment can be considered a better explanation of the data at hand because it captures an element of homologous similarity between the sequences of *A* and *B* that is not retained in the second one. However the tree of the first tree alignment is rooted, *A* and *B* share the absence of bases 4–7 with their direct ancestor. Depending on the position of the root, these three contiguous nodes lack the insertion of that subsequence, or they share its deletion; in both cases, this comes down to one unit similarity that can be explained as a homology. On the second tree alignment, the shared absence of bases 4–7 in *A* and *B* must be explained as a homoplasy. The main conclusion that can be drawn from this simple example is that sequence homology has a component that cannot be reduced to mere base-to-base composition. This component I shall refer to as *homology of subsequences*, as opposed to *base-to-base* or *compositional homology* within homologous subsequences.

The two components of sequence homology can be optimized separately but there would be little use in doing so. When just optimizing base-to-base similarities, gaps will be inserted 'at will' to maximize matches (Smith *et al*. 1981, p. 42). On the other hand, maximizing subsequence homology without regard for the composition of those subsequences comes down to optimizing the length of the observed sequences as a regular unordered character, irrespective of the amount of substitutions that are implied. Optimized in isolation, neither will in general result in a globally optimal explanation of the data.

Instead, what is needed is an optimal balance between subsequence and compositional homology. This optimal balance can be found by using a cost regime that is the sum of the two cost regimes that are involved, provided that there is a mechanism to avoid or deal with logical contradictions between optimizations of both components. Such a mechanism is implicit in tree alignments because tree alignments are internally consistent explanations of the data. Therefore, expressions to describe the amount of subsequence homology and the amount of compositional homology in tree alignments can be derived independently and then simply summed to get an expression for the total amount of sequence homology. This expression, finally, can be used for purposes of optimization.

Quantifying the amount of subsequence homology of a tree alignment

The amount of subsequence similarity in a tree alignment that can be interpreted as homology can be measured indirectly and in a relative way by counting n_{indels}, the number of independent indel events, provided that the insertion/deletion of a series of contiguous bases is counted as a single event. This is so because each such indel event effectively marks a subsequence that is not homologous across a branch. Therefore, an independent indel event can be seen as an independent unit of non-homology in subsequence homology.

As discussed above, the cost of a tree alignment is obtained as the sum of the costs of the pairwise alignments across the branches of the tree. Technically, counting independent indel events in such a pairwise alignment is achieved by setting substitution costs to 0, gap opening cost to 1, and gap extension cost to 0. In addition, when evaluating such a pairwise alignment, paired gaps have to be removed first, a procedure that Altschul (1989) called *projection*. Projection is required because paired gaps just indicate that both sequences miss something that is present elsewhere on the tree and because the indel events that caused such a shared absence are accounted for along other branches. As an example, going from *-gaat---ccct-* to *-gaat--ccccc-* in, for example, the second tree alignment of Fig. 6.14, (see below) means going from *gaat-ccct* to *gaatccccc*. As far as subsequence homology is concerned, this comes at cost 1 (1 times the gap opening cost of 1 plus 0 times the extension cost of 0).

Quantifying the amount of compositional homology of a tree alignment

Specifying an expression for compositional similarity that can be explained as homology is more elaborate. A tree alignment can be seen as a regular multiple alignment with, for each position, reconstructions at the inner nodes. If, in a single column, the tree path between two observed bases passes through an inner node that is optimized as a unit gap character, these bases are not comparable because they are part of non-homologous subsequences; if, on the other hand, the connecting path has no nodes with unit gaps, they belong to homologous subsequences; more specifically, they occur at the same position within those homologues. I refer to such bases as *comparable bases*.

The observed bases in a single column of a tree alignment can be sorted into a number of groups such that two bases from the same group are comparable but two bases from different groups are not comparable. I shall refer to these groups of comparable bases as *subcharacters*, a concept that is closely related to the concept of regions as defined above, and denote the number of subcharacters in a column of a tree alignment as nsc_c. This number is related but not identical to the number of indel events in which this column of the alignment is involved.

Within a subcharacter, denote the number of observed bases as nob_{sc}. If two such bases are identical and all nodes in the path that connects them are labeled with that same base, then the two bases match and their shared presence can be explained as a homology. If any node in the path that connects two such identical bases has a base that is different, then they don't match and their shared presence cannot be explained as a homology. Two non-identical bases of a subcharacter or two bases that belong to different subcharacters, finally, do not contribute to base-to-base homology. The minimum number of pairwise comparisons that have to be made to classify the bases of a subcharacter into subgroups of such matching bases is $nob_{sc} - 1$. The number of mismatches nmm_{sc} in any such set of $nob_{sc} - 1$ independent pairwise comparisons can be thought of as the number of base substitutions or steps within the subcharacter.

With these definitions, the amount of compositional homology in a subcharacter is obtained just as the amount of homology in a regular character: the maximum number of independent pairwise comparisons minus its number of steps, or $nob_{sc} - 1 - nmm_{sc}$. With nob_c the total number of observed bases and nmm_c the total number of substitutions in a column of a tree alignment, the amount of compositional homology in a column is $nob_c - nsc_c - nmm_c$. The amount of compositional homology in the whole tree alignment is the sum of this value over all columns. Switching signs, $nsc_c + nmm_c - nob_c$ describes a cost function that varies directly with compositional homology in a column. In this expression, nsc_c can be considered a cost factor that accounts for local loss of compositional homology due to indel events (that may encompass multiple neigbouring columns), and nmm_c a regular substitution cost factor.

Maximizing homology in sequence characters

Adding it up, the total amount of homology of different tree alignments for a given set of sequences can be compared using cost function $n_{indels} + \Sigma(nsc_c + nmm_c - nob_c)$, where the summation is over all columns of the tree alignment: the lower the cost, the higher the amount of homology. In this expression, losses in subsequence homology and compositional homology are weighted equally. Differential weighting, for example to downweight subsequence homology, can be done by applying different weights to the two terms that are involved. As Σnob_c is identical for different tree alignments for the same data, the cost function for a tree alignment can be reduced to $n_{indels} + \Sigma(nsc_c + nmm_c)$. Using n_{subc} for Σnsc_c and n_{subst} for Σnmm_c, the relative amounts of total homology of two different tree alignments can be compared using $n_{indels} + n_{subc} + n_{subst}$, the sum of indel events, subcharacters, and substitutions.

Alternatively, the problem can be presented as a maximization of a similarity measure; this similarity measure would count independent homologous base-to-base matches but assign a penalty to indel events, much as the original algorithm of Needleman and Wunsch (1970). More specifically, the penalty would be -1 for each indel event in the tree alignment, irrespective of the length of the indel. In comparisons of two and three sequences, such similarity measures with length independent gap penalties have been studied by Fredman (1984) (*fide* Hein 1989a, p. 650).

In Figs. 6.11–6.15, the positions of all inferred indel events are indicated throughout the tree alignments, using vertical bars. The subsequences that are defined in that way can be considered *logical subsequences*. In simple cases, such logical subsequences are identical to the subsequences that effectively take part in the inferred indel events (e.g. Figs 6.11–6.14), but in more complex cases a single inferred indel event along a particular branch can affect a series of contiguous logical subsequences (see Fig. 6.15 for examples). The total number of subcharacters in a tree alignment can be easily determined as the sum of the lengths of its different homologous logical subsequences.

For any given tree alignment, $n_{indels} + n_{subc} + n_{subst}$ is a straightforward expression that is easily checked, but finding the tree alignment(s) for which this expression is minimal is quite something else. Even for a single given tree, the problem of deciding if a tree alignment is optimal has been shown to be NP-complete (Wang and Jiang 1994). Algorithmically, as the subsequence homology component requires use of variable gap costs (gap opening cost 1, gap extension cost 0), the algorithms of Sankoff (1975) and Sankoff and Cedergren (1983) are not adequate. Altschul (1989) does accomodate variable gap costs but still this is not sufficient because his algorithm does not keep track of the number of subcharacters in a column. This directly implies that the current cost function cannot be expressed just in terms of substitution, gap opening, and gap extension costs. To optimize this function, the dynamic programming recurrences of Altschul (1989) would have to be adapted and extended to keep track of observed bases and subcharacters in columns as well.

6.3.4 Discussion

So, when applied to sequence data, the simple principle of maximizing similarity that can interpreted as homology, in a logically correct way, leads to a preference for those trees on which the sum of indel events, base substutions, and subcharacters is minimal. In this final subsection, some properties and wider connections of this parsimony criterion are discussed.

Heuristics

Even with simple Hamming distances, as when using Fitch (1971) optimization of prior alignments, the problem of deciding if a tree is optimal is NP-complete (Foulds and Graham 1982). So, when combining tree search and tree alignment, one NP-complete problem is nested within another. As pointed out by Hein (1989a, p. 651), the computational complexity of this problem makes the use of heuristic approximations unavoidable. Examples of algorithms for heuristic approximations of optimal tree alignment costs, or algorithms that can be interpreted as such, can be found in, for example, Sankoff *et al*. (1973, 1976), Hein (1989a, b), Jiang and Lawler (1994), Wang *et al*. (1996), Wheeler (1996, 1999, 2003c; all available in Wheeler *et al*. 2003, where they are tightly integrated with a wide range of tree search heuristics; see De Laet and Wheeler 2003), and Schwikowski and Vingron (1997, 2003). Still other approaches can be found in the reviews of Vingron (1999) and Notredame (2002).

It currently remains largely an open question how well these various approaches perform in practice. In the end, even the use of an *a priori* alignment can be seen as a quick and dirty heuristic for the analysis of a sequence character. Even if any single such analysis is too shallow to be satisfactory, analyses of many different prior alignments may be effectively combined into a more elaborate search strategy, following the heuristic logic as developed in Farris *et al*. (1996) (see also Goloboff and Farris 2001).

Most heuristic tree alignment methods attack the optimal tree alignment problem by approximate decomposition into a set of simpler problems that can easily be solved exactly using pairwise alignments (e.g. Hein 1989a; Wang *et al*. 1996; Wheeler 1996, 1999) or threewise alignments on a star tree (e.g. Sankoff *et al*. 1973; Wheeler 2003c). Interestingly, compositional homology in a pairwise alignment amounts to the number of base matches, a number that can be maximized by setting the unit gap cost to half the substitution cost (Smith *et al*. 1981). To maximize total sequence homology in a pairwise alignment, an additional penalty has to be added for losses in subsequence homology, which, as discussed above, can be done using the gap opening cost. With equal weighting of both components of homology, this penalty equals the substitution cost. As an example, using a substitution cost of 2, the corresponding gap-opening cost is $2+1$, and the corresponding gap extension cost 1. The same result holds for threewise comparisons on a star tree.

Beyond three sequences this simpler cost regime is no longer equivalent to the criterion developed here, as can be seen from the following counterexample. The tree alignment of Fig. 6.11b explains the sequence character of Fig. 6.11a better than Fig. 6.11c because it can explain an additional independent pairwise base match: the *a* that terminates the sequences of *B* and *D*. This difference is correctly measured by the sum of indels, subcharacters, and substitutions, but with the simpler cost regime, both tree alignments come at the same cost of 12. In more complex examples, such situations can lead to a preference for different trees alltogether. The simpler cost regime may nevertheless be a good choice when using heuristic tree alignment methods that are based on pairwise or threewise comparisons of sequences.

For some approximation methods an upper bound can be established for their deviation of optimality. As an example, consider lifted alignments (Jiang and Lawler 1994; Wang *et al*. 1996; see also Wheeler 1999; Lutzoni *et al*. 2000), in which possible inner-node sequences are chosen from and restricted to the set of observed sequences. Under these restricted conditions, an efficient algorithm exists to find the optimal assignments of sequences to inner nodes of a given tree, and the resulting tree alignment can be shown to have a cost that is at most twice the cost of the unrestricted

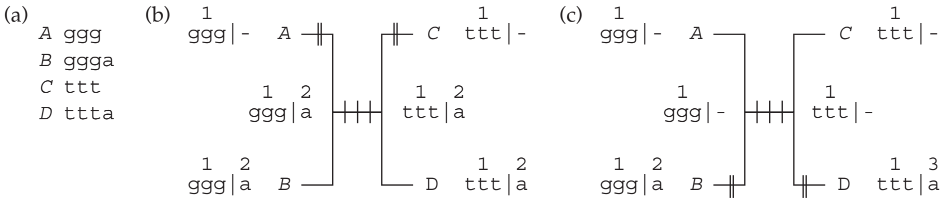

Figure 6.11 An example of the parsimony criterion for sequence characters. (a) A sequence character. (b) An optimal tree alignment on the optimal tree. (c) A suboptimal tree alignment on the optimal tree (same number of indel events and substitutions, but one more subcharacter). Single bars across branches indicate substitutions, double bars indel events. Logical subsequences are indicated using vertical bars, and numbered for clarity.

optimum for that tree (Wang *et al*. 1996) As discussed by Gusfield (1997, p. 358), such bounded-error approximation methods can help to understand the behaviour of difficult optimization problems; from a practical point of view, they may be combined with other methods, such as local improvement methods, to obtain more elaborate heuristic search strategies.

Inapplicables

The example of Fig. 6.12 illustrates that indel events divide the sequences of the tree alignment into subsequences that can be considered independently: the two optimal alignments that are shown have identical subsequences and only differ in the way that those subsequences (and their subcharacters) are presented. Incidentally, this example also shows that postulated indel events may improve the explanation of the data even in cases where all observed sequences have the same length.

This independence is a direct consequence of the fact that, in the current approach, base-to-base comparisons are only made within subsequences that can be explained as homologs. As a consequence, comparisons of sequences and their bases automatically occur at the correct levels of generality, and the problems with inapplicables that Maddison (1993) described simply dissolve. Indeed, Maddison (1993, p. 580) observed that all solutions that he considered to deal with inapplicables were in the end problematic because they did not properly restrict counting of steps to parts of trees where comparisons were valid, and he correctly surmised that an eventual solution would lie in the development of new algorithms. Most cases of inapplicability, however, would not require an algorithm as complex as the one discussed here, because there are fewer degrees of freedom in *a priori* acceptable hypotheses of homology.

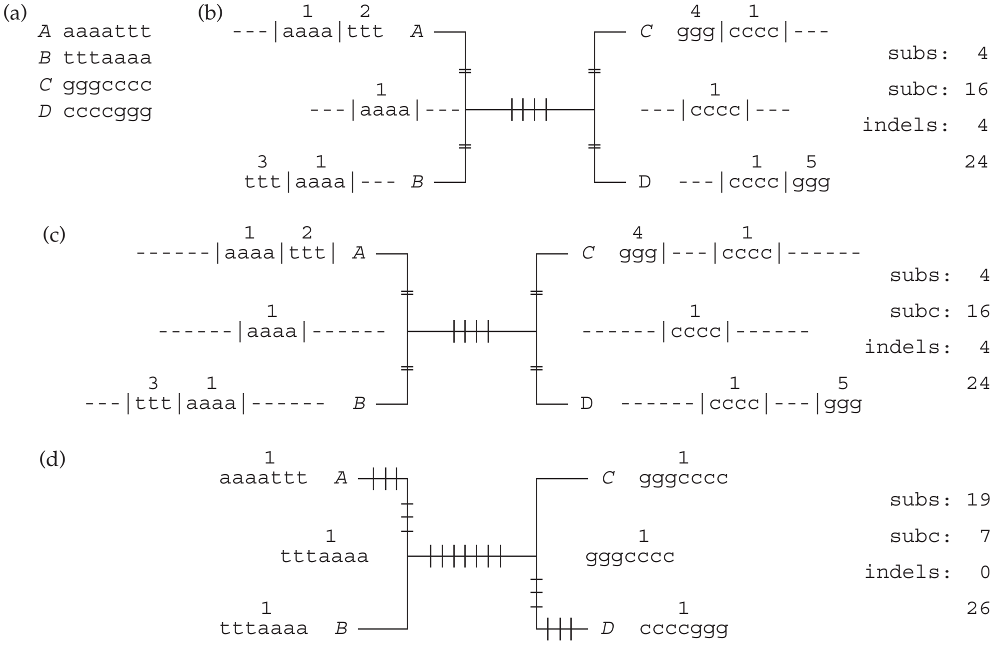

Figure 6.12 An example of the parsimony criterion for sequence characters. (a) A sequence character in which all sequences have equal length. (b, c, d) Three tree alignments of the character on the optimal tree (*A B*)(*C D*). The first two, requiring four indel events, are optimal; the third, not requiring indel events, is suboptimal by two units. The two optimal alignments that are shown imply the same five subsequences that take part in indel events and differ only in the way that these subsequences are presented (still other possibilities exist). Subs, subc, and indels are numbers of substitutions, subcharacters, and indel events. Single bars across branches indicate substitutions, double bars indel events. Logical subsequences are indicated using vertical bars, and numbered for clarity.

Consider again the multiple alignment of Fig. 6.8b, but now assume that the four columns are regular independent single-column characters, with a dash indicating inapplicability. Obviously, in this case there is no need to examine alternative groupings of states, such as in Fig. 6.8c, during tree search and optimization. Permitting such shifts would lead to the same problems as when using absence/presence coding of individual states. As the computational complexity of the current approach mostly derives from the need to examine alternative groupings of bases when optimizing sequences on a tree, this restriction has as a fortunate consequence that the general algorithm for dealing with this kind of inapplicability is much simpler and faster (De Laet 2003).

Maximizing homologous similarity vs. mimimizing transformations

The parsimony criterion as discussed here relies on the notion that one indel event counts as one unit loss of subsequence homology, irrespective of the number of bases that are involved. But this does not mean that it would in general produce trivial alignments that are obtained by simply juxtaposing all observed sequences, which requires only as many insertion events as there are sequences. An example is presented in Fig. 6.13. In the optimal tree alignment of Fig. 6.13b, two independent pairwise base matches can be explained as homology. The trivial alignment that is obtained by juxtaposing all observed sequences (Fig. 6.13c) has no such base matches. In addition, compared to the first tree alignment, it has has four independent instances of subsequence non-homology. The total difference in explanatory power thus equals six, which is reflected in the relative tree scores.

This shows that the current criterion is not a minimum evolution method: the second tree alignment of Fig. 6.13 requires only four mutations (four insertions of subsequences of length four) but it is considered a much worse explanation of the data than the first one, which requires 10 mutations (10 substitutions). Given that one of the terms in the minimization for sequence character homology is the number of subcharacters, a quantity that has no direct relationship with evolutionary transformations, the non-equivalence of both approaches when dealing with sequence characters should come as no surprise. But this non-equivalence with minimization of evolutionary transformations does not imply that the current method is not logically capable of phylogenetic interpretation. Such an interpretation, however, is in terms of unit statements of similarity that can be explained in a logically consistent way as identity through

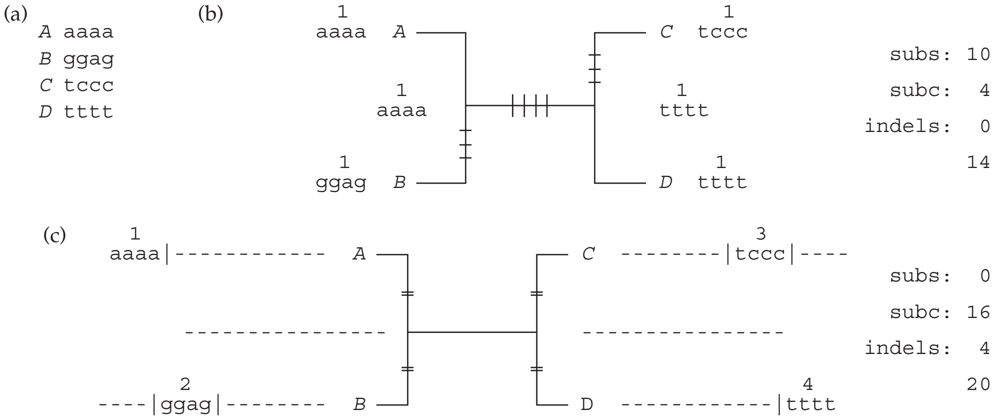

Figure 6.13 An example of the parsimony criterion for sequence characters. (a) A sequence character. (b, c) Two tree alignments on the optimal tree (*A B*)(*C D*). The first is optimal. The second, obtained by simply juxtaposing all observed sequences, is suboptimal by six units. Subs, subc, and indels are numbers of substitutions, subcharacters, and indel events. Single bars across branches indicate substitutions, double bars indel events. Logical subsequences are indicated using vertical bars, and numbered for clarity.

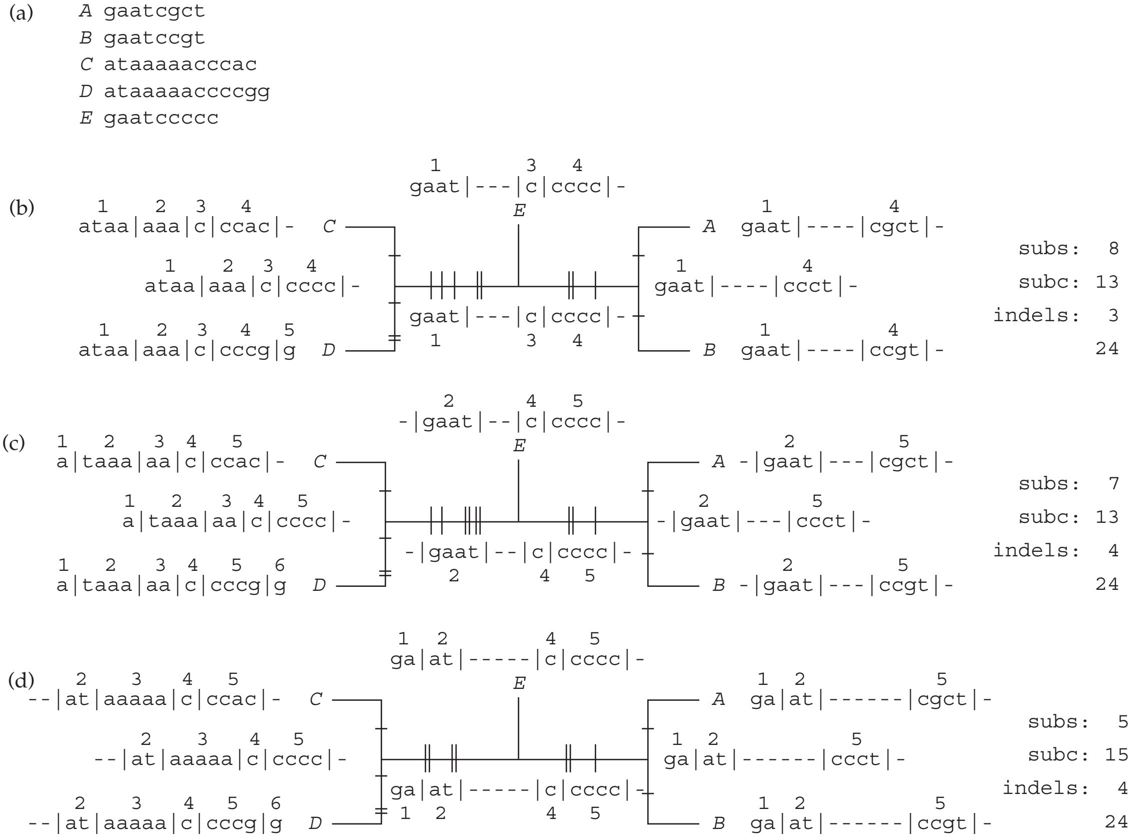

Figure 6.14 An example of the parsimony criterion for sequence characters. (a) A sequence character. (b, c, d) Three optimal tree alignments on its optimal tree. Subs, subc, and indels are numbers of substitutions, subcharacters, and indel events. Single bars across branches indicate substitutions, double bars indel events. Logical subsequences are indicated using vertical bars, and numbered for clarity.

common descent and inheritance, and not in terms of numbers of transformations that are required to that effect.

An example where different optimal tree alignments on the best tree have different numbers of indels plus substitutions is presented in Fig. 6.14. The two first tree alignments have more indel events plus substitutions than the third one (11 versus 9), but despite this higher total number of mutations, they provide an equally good overall explanation of the data in terms of the amount of total sequence similarity that can be explained as homology. More precisely, the first alignment accomodates 29 independent pairwise matches among observed bases, the second 30, and the third one 30 as well, as easily verified by examining the tree alignments column by column. So just considering compositional homology, the first explanation is suboptimal. The difference, however, is exactly offset by its lower loss in subsequence homology (three indels versus four and four). With the cost regime that is advocated by Frost *et al*. (2001) (all costs equal), the optimization of Fig. 6.14c is preferred (cost 12 vs. costs 13 for 14b and 14 for 14d).

The difference between both cost regimes is further illustrated in Fig. 6.15. Maximizing the amount of sequence similarity that can be interpreted as homology, the tree of Fig. 6.15b is optimal, and an optimal tree alignment is shown. The tree of Fig. 6.15c is suboptimal by two units, as can be seen from the optimal alignment that

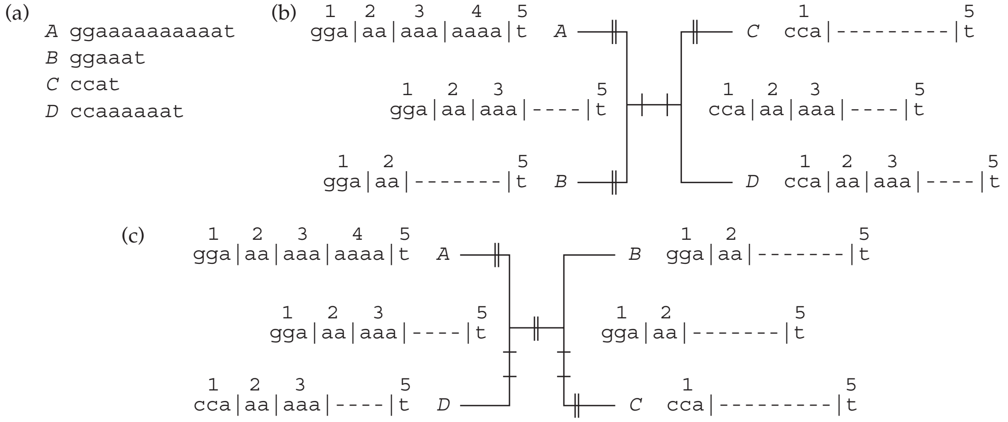

Figure 6.15 An example of the parsimony criterion for sequence characters. (a) A sequence character. (b) An optimal tree alignment on the optimal tree. (c) An optimal tree alignment on a suboptimal tree. Single bars across indicate substitutions, double bars indel events. Logical subsequences are indicated using vertical bars, and numbered for clarity. The number of subcharacters in both optimizations is the same.

is shown. Under the costs of Frost *et al*. (2001) the tree alignments of Figs 6.15b and 6.15c are also optimal for their respective trees, but the ranking of the trees reverses: the second tree is now preferred (costs 14 vs. 13). This shift in preference is a consequence of counting an indel event of length *k* as *k* events, as implicitly advocated by Frost *et al*. (2001). In this example, this amounts operationally to treating the lengths of the gaps that are involved as an ordered character.

A more extreme example of the same phenomenon occurs with a sequence character such as *ttaatt*, *ttaaatt*, *ttaaaatt*, and *ttaaaaatt* for terminals *A*, *B*, *C*, and *D*. With the cost regime of Frost *et al*. (2001), unrooted tree (*A B*)(*C D*) is preferred because, operationally, it best groups the series of *a*'s in the middle of the observed sequences according to their length. With the cost regime that maximizes homology, the three different unrooted trees for four terminals are considered equally good explanations of the character.

The preference of Frost *et al*. (2001, pp. 354–355) for equal substitution and unit gap costs follows from their position that all hypothesized evolutionary transformations should be weighted equally. However, this cost regime only accomplishes such equal weighting under the very restrictive assumption that indels only affect single bases, which constitutes a severe knowledge claim about the processes that shape sequence evolution. It is hard to see then how this approach 'maximizes the explanatory power of all lines of evidence' (Frost *et al*. 2001, p. 354) even more so if one considers their apparent position that methods that make severe knowledge claims can be safely ignored (Frost *et al*. 2001, p. 354). No comparable claim is present in the current method, in which the lengths and positions of subsequences that take part in indel events are left open to optimization.

A similar methodological asymmetry exists between methods that impose irreversibility of inferred character evolution and methods that leave the possibility of reversal open during phylogenetic analysis. An extensive discussion of the issues that are involved can be found in Farris (1983, pp. 24–27). Frost *et al*. (2001) did not discuss such issues. In fact, they did not not even provide arguments why equal weighting of all evolutionary transformations should lead to equal substitution and unit gap costs. It can reasonably be argued that the principle of equal weighting of all transformations is instead better implemented by using equal substitution and gap costs, irrespective of the length of the gaps that are involved. However, for most sequence characters this cost regime would lead to trivial alignments such as in Fig. 6.13c, requiring only as many transformations as there are terminals, irrespective of the tree that is considered. Again, it is hard to see how such optimizations can be considered to maximize explanatory power. Yet they are optimal under the notion of minimizing equally weighted transformations.

Sequence characters and branch support

The example of Fig. 6.13 illustrates an interesting consequence for the concept of branch support. Consider the tree alignment of Fig. 6.13b. In that alignment, the (*A B*)(*C D*) branch is supported, not because of the four substitutions on that branch, but because collapse of the branch—resulting in an unresolved tree—would remove either the *a*–*a* base match between *A* and *B* or the *t*–*t* base match between *C* and *D*. This is in line with the observation of Farris *et al*. (2001a) that branch lengths do not measure support. Instead, support for any single branch is measured as the degree to which removal of the branch worsens the explanation of the data, which holds for sequence and non-sequence data alike. This, by definition, is Bremer (1988) support.

Alternatively, one could measure robustness of a branch using the jackknife (Farris *et al*. 1996) or related methods. However, as sequence characters have no predefined single-column characters, pseudoreplicates cannot be constructed in the usual way. This problem can be solved by resampling at the level of individual bases in the sequences to be compared, such that unsampled bases are made uninformative with a probability equal to the character removal probability of regular jackknifing (operationally, this can be done by replacing a base with a polymorphism code for '*a* or *c* or *g* or *t* or -'; or, a bit more conservative, for '*a* or *c* or *g* or *t*'). With a removal probability of 0.37, the (*A B*)(*C D*) branch in the above example would not survive, as it depends on the simultaneous presence of the four bases mentioned above. With the conservative approach, the probability that all four are retained in a pseudoreplicate is only $(1 - 0.37)^4$.

A likelihood conjecture

Miklós *et al*. (2004) recently described a probabilistic model of sequence evolution that allows insertions and deletions of arbitrary length, a more general approach than Thorne *et al*. (1992), the first probabilistic method that incorporated indels that affect multiple residues at once. In their model, substitutions are described using a regular time-reversible rate matrix; indels are modelled such that the rates for insertions as a function of their length *k* are a geometric function of *k*, and such that the ratio between the rates of insertions and deletions of length *k* is a constant.

Miklós *et al*. (2004) only dealt with comparisons of two sequences, but the model can in principle be extended to simultaneous comparison of more than two sequences that are related by a binary tree, similarly as Hein (2001) extended the two-sequence model of Thorne *et al*. (1991), the first stochastic model to include insertions and deletions (single residue indels only). In the approach of Hein (2001), rate parameters are assumed to be constant throughout the sequences. Removal of assumptions of that kind would turn the model into a no-common-mechanism model akin to the model of Tuffley and Steel (1997, pp. 584, 597) for regular *r*-state characters.

Envisioning such a double extension of the model of Miklós *et al*. (2004) it can be conjectured that, under a wide range of possible non-fixed rates, the trees that are found with a parsimony criterion along the lines as described here are also trees of maximum likelihood. As with single-column characters (see above), this does not imply that such a probabilistic process model would exhaustively describe and capture the current method.

Beyond sequence characters: the genome

Most examples above consist of data sets with just a single sequence character, but data sets can have several such characters, and in addition any number of single-column characters. Exactly which observations are coded as characters, the subject of character analysis, is ultimately outside the realm of the technical aspects of further analysis that have been discussed in this section. For sequence characters, a widely used criterion for establishing hypotheses of putative sequence homology is almost identical to the technology to obtain those sequences in the first place: whatever is amplified using a particular primer pair. In addition, various other criteria can be used to identify biologically relevant structures, such as exons and introns in protein coding sequences, or stems and loops in rRNA sequences (see, e.g., Kjer 1995; Giribet 2002).

On the basis of such criteria, even contiguous stretches of the genome can be subdivided into sequences of sequence characters that can be optimized separately. When doing so, it may be a legitimate concern that the subsequent analysis might be constrained and even biased by preconceived ideas about the evolution of such structures. However, given that the complexity of the calculations when dealing with sequence characters makes the use of heuristics and approximations unavoidable, the procedure of breaking up long sequences in smaller components prior to analysis may very well be part of a heuristic search strategy. This approach could be especially powerful when combined with heuristic multiple alignment methods that try to assemble global alignments from alignments of fragments that are dynamically identified (e.g. Morgenstern *et al*. 1996; Morgenstern 2004).

On a more fundamental level, sequence characters as discussed here are thought to be hierarchically related through indels and substitutions only. This may be a biologically plausible assumption for shorter parts of the genome, but it definitely breaks down for complete genomes, where other processes such as inversions, duplications, and translocations play a role as well. Over the past few years, many combinatorial algorithms have been developed to study such phenomena (see, e.g., Sankoff and Nadeau 2000), and heuristic multiple-alignment methods that incorporate such rearrangment events are becoming available (see, e.g., Brudno *et al*. 2003, 2004). It remains an open question how such methods can be interpreted or generalized to accomodate a parsimony criterion as developed here.

Such extensions may well lead to revisions or further elaborations of the current framework. Consider, for example, a process such as lateral transfer, which may well play an important role in the evolution of genomes (see, e.g., Kunin and Ouzounis 2003), or speciation through allopolyploidization (see, e.g., Vander Stappen *et al*. 2002). For any data set, positing sufficient such events in any phylogenetic tree will permit to explain all observed similarities as historically identical, whether through regular ancestor–descendant relationships of organisms or through non-hierarchic processes such as lateral transfer. It may be sufficient to restrict the current criterion to the former case, but, alternatively or additionally, a more general criterion might be conceived that maximizes the difference between similarity that can be explained as historical identity, whatever the underlying processes, and the minimum number of hypothesized historical events required to that effect.

This second approach would need careful elaboration of a broader theoretical concept of explanation than used here, which is beyond the scope of this chapter. However, one way to go would be to couple the principle of maximizing conformity between observation and theory to the principle of choosing the simplest theory or theories that can explain the data, which would lead to a true synthesis of two different but interwoven lines of argument that can be found in the work of Farris (see, e.g., Farris 1982b, 1983). As discussed extensively in this paper, the first principle leads to maximization of similarity that can be explained as homology. The second principle requires a measure of the simplicity of a phylogenetic explanation, which may well be the minimum number of logically distinct historical events that have to be postulated. The rationale for a combined optimality function as above would then be to find an optimal balance between both principles.

For single-column character data and under the above restriction, that approach would operationally be equivalent to the current parsimony criterion, because in such cases it amounts to minimizing twice the amount of homoplasy. For sequence characters as defined here (only indels and substitutions), it would amount to minimizing

$2n_{indels} + n_{subc} + 2n_{subst}$, which would obviously change details of several examples discussed in this section. For example, both trees of Fig. 6.13 are then considered equally good explanations; or the two first trees of Fig. 6.14 become suboptimal by two units. But the main conclusions, and especially those based on data symmetries, would remain valid.

6.4 Acknowledgments

This chapter was prepared on a gnu/linux system, mostly using vim and LaTeX. Thanks to Victor Albert, Norberto Giannini, Pablo Goloboff, Mark Simmons, Hubert Turner, John Wenzel, and two anonymous reviewers for constructive criticisms. Thanks to Kevin Nixon for bandwidth and disk space to make the computer program *goechel* available at www.cladistics.com. Goechel is a java shell around POY (Wheeler *et al*. 2003) to perform jackknife analyses as discussed in 6.3.4 (subsection *Sequence characters and branch support*). The author holds a return grant from The Belgian Federal Science Policy Office. Lastly I wish to mention Steve Farris. It's been my privilege to have him as a member of my PhD committee, late last century. Ever since, I consider myself a student of his, as should be obvious by even just a cursory glance at this paper. Thanks for it all, Steve!

References for: *De Laet, J. 2005. Parsimony and the problem of inapplicables in sequence data. Pp. 81-116 in Albert, V.A. (ed.) Parsimony, phylogeny and genomics. Oxford University Press, ISBN 0-19-856493-7.*

Altschul, S. F. 1989. Gap costs for multiple alignment. J. Theoret. Biol. 138: 297--309.

Boyd, R. 1991. Confirmation, semantics, and the interpretation of scientific theories. In The philosophy of science (eds. R. Boyd, P. Gasper, and J. D. Trout), pp. 3--35. MIT Press, Cambridge, Massachusetts.

Bremer, K. 1988. The limits of amino acid sequence data in angiosperm phylogenetic reconstruction. Evolution 42: 795--803.

Brudno, M., Chapman, M., Göttgens, B., Batzoglou, S., and Morgenstern, B. 2003. Fast and sensitive multiple alignment of large genomic sequences. BMC Bioinformatics 4: 66.

Brudno, M., Poliakov, A., Salamov, A., Cooper, G. M., Sidow, A., Rubin, E. M., Solovyev, V., Batzoglou, S., and Dubchak, I. 2004. Automated whole-genome multiple alignment of rat, mouse, and human. Genome Research 14: 685--692.

Carillo, H., and Lipman, D. 1988. The multiple sequence alignment problem in biology. SIAM J. Appl. Math. 48: 1073--1082.

Caroll, L. 1872. Through the looking glass. Macmillan, London.

Carpenter, J. M. 2003. On "Molecular phylogeny of Vespidae (Hymenoptera) and the evolution of sociality in wasps". American Museum Novitates 3389: 1-20.

De Laet, J. 1997. A reconsideration of three-item analysis, the use of implied weights in cladistics, and a practical application in Gentianaceae. PhD thesis, University of Leuven.

De Laet, J. 2003. Parsimony algorithms for characters that are inapplicable in some terminals (Abstract, 21st annual meeting of the Willi Hennig Society, Helsinki 2002). Cladistics 19: 148.

De Laet, J. 2004. When one and one is not two: parsimony analysis of sequence data (Abstract, 22nd annual meeting of the Willi Hennig Society, New York 2003). Cladistics 20: 81.

De Laet, J., and Smets, E. 1998. On the three-taxon approach to parsimony analysis. Cladistics 14: 363--381.

De Laet, J., and Wheeler, W. 2003. POY version 3.0.11. (Wheeler, Gladstein and De Laet, May 6 2003). Command line documentation. Available from the first author and at ftp://ftp.amnh.org/pub/molecular/poy.

de Pinna, M. C. C. 1991. Concepts and tests of homology in the cladistic paradigm. Cladistics 7: 367--394.

Endress, P. K. 1994. Diversity and evolutionary biology of tropical flowers. Cambridge University Press, Cambridge.

Farris, J. S. 1970. Methods for computing Wagner trees. Syst. Zool. 19: 83--92.

Farris, J. S. 1972. Estimating phylogenetic trees from distance matrices. American Naturalist 106: 645--668.

Farris, J. S. 1978a. Inferring phylogenetic trees from chromosome inversion data. Syst. Zool. 27: 275--284.

Farris, J. S. 1979. The information content of the phylogenetic system. Syst. Zool. 28: 483--519.

Farris, J. S. 1982a. Outgroups and parsimony. Syst. Zool. 31: 328--334.

Farris, J. S. 1982b. Simplicity and informativeness in systematics and phylogeny. Syst. Zool. 31:

413--444.

Farris, J. S. 1983. The logical basis of phylogenetic analysis. In Advances in Cladistics Vol. 2 (eds. N. I. Platnick, and V. A. Funk), pp. 7--36. Columbia University Press, New York, New York.

Farris, J. S. 1986a. On the boundaries of phylogenetic systematics. Cladistics 2: 14--27.

Farris, J. S. 1989. The retention index and the rescaled consistency index. Cladistics 5: 417--419.

Farris, J. S. 1991. Hennig defined paraphyly. Cladistics 7: 297--304.

Farris, J. S. 1997. Cycles. Cladistics 13: 131--144.

Farris, J. S. 1999. Likelihood and inconsistency. Cladistics 15: 199--204.

Farris, J. S., and Kluge, A. G. 1986. Synapomorphy, parsimony, and evidence. Taxon 35: 298--315.

Farris, J. S., Källersjö, M., Albert, V. A., Allard, M., Anderberg, A., Bowditch, B., Bult, C., Carpenter, J. M., Crowe, T. M., De Laet, J. et al. 1995. Explanation. Cladistics 11: 211--218.

Farris, J. S., Albert, V. A. A., Källersjö, M., Lipscomb, D., and Kluge, A. G. 1996. Parsimony jackknifing outperforms neighbor-joining. Cladistics 12: 99--124.

Farris, J. S., Källersjö, M., and De Laet, J. E. 2001a. Branch lengths do not indicate support - even in maximum likelihood. Cladistics 17: 298--299.

Farris, J. S., Kluge, A. G., and De Laet, J. E. 2001b. Taxic revisions. Cladistics 17: 79--103.

Felsenstein, J. 1978a. Cases in which parsimony or compatibility methods will be positively misleading. Syst. Zool. 27: 401--410.

Felsenstein, J. 1978b. The number of evolutionary trees. Syst. Zool. 27: 27--33.

Felsenstein, J. 1979. Alternative methods of phylogenetic inference and their interrelationships. Syst. Zool. 28: 49--62.

Felsenstein, J. 1981a. Evolutionary trees from DNA sequences: a maximum likelihood approach. J. Mol. Biol. 17: 368--376.

Felsenstein, J. 2004. Inferring phylogenies. Sinauer Associates, Sunderland, Massachusetts.

Feng, D., and Doolittle, R. F. 1987. Progressive sequence alignment as a prerequisite to correct phylogenetic trees. J. Mol. Biol. 25: 351--360.

Fitch, W. M. 1970. Distinguishing homologous from analogous proteins. Syst. Zool. 19: 99-113.

Fitch, W. M. 1971. Toward defining the course of evolution: minimal change for a specific tree topology. Syst. Zool. 20: 406--416.

Foulds, L. R., and Graham, R. L. 1982. The Steiner problem in phylogeny is NP-complete. Adv. Appl. Math. 3: 43--49.

Fredman, M. L. 1984. Algorithms for computing evolutionary similarity measures with length independent gap penalties. Bull. Math. Biol. 46: 545--563.

Freudenstein, J. V., Pickett, K. M., Simmons, M. P., and Wenzel, J. W. 2003. From basepairs to birdsongs: phylogenetic data in the age of genomics. Cladistics 19: 333--347.

Frost, D. R., Rodrigues, M. T., Grant, T., and Titus, T. A. 2001. Phylogenetics of the lizard genus Tropidurus (Squamata: Tropiduridae: Tropidurinae): direct optimization, descriptive efficiency, and sensitivity analysis of congruence between molecular data and morphology. Mol. Phylogenet. Evol. 21: 352-371.

Giribet, G. 2002. Relationships among metazoan phyla as inferred from 18S rRNA sequence data: a methodological approach. In Molecular systematics and evolution: theory and practice (eds. R. DeSalle, G. Giribet, and W. Wheeler), pp. 85--101. Birkhäuser Verlag, Basel.

Giribet, G., and Wheeler, G. 1999. On gaps. Mol. Phylogenet. Evol. 13: 132--143.

Goloboff, P. A. 1996b. Methods for faster parsimony analysis. Cladistics 12: 199--220.

Goloboff, P. A. 1999. Analyzing large datasets in reasonable times: solutions for composite optima. Cladistics 15: 415--428.

Goloboff, P. A. 2003. Parsimony, likelihood, and simplicity. Cladistics 19: 91--103.

Goloboff, P. A., and Farris, J. S. 2001. Methods for quick consensus estimation. Cladistics 17: S26--S34.

Grant, T., and Kluge, A. G. 2004. Transformation series as an ideographic character concept. Cladistics 20: 23--31.

Gusfield, D. 1997. Algorithms on strings, trees, and sequences. Cambridge University Press, Cambridge.

Hartigan, J. A. 1973. Minimum mutation fits to a given tree. Biometrics 29: 53--65.

Hein, J. 1989a. A new method that simultaneously aligns and reconstructs ancestral sequences for any number of homologous sequences, when the phylogeny is given. Mol. Biol. Evol. 6: 649--668.

Hein, J. 1989b. A tree reconstruction method that is economical in the number of pairwise comparisons used. Mol. Biol. Evol. 6: 669--684.

Hein, J. J. 2001. An algorithm for statistical alignment of sequences related by a binary tree. In Pacific Symposium on Biocomputing 2001 (eds. R. B. Altman, A. K. Dunker, L. Hunter, K. Lauderdale, and T. E. Klein), volume 6, pp. 179--190. World Scientific, Signapore.

Hendy, M. D., and Penny, D. 1982. Branch and bound algorithms to determine minimal evolutionary trees. Math. Biosc. 59: 277--290.

Hennig, W. 1950. Grundzüge einer Theorie der phylogenetischen Systematik. Deutscher Zentralverlag, Berlin.

Hennig, W. 1966. Phylogenetic systematics. University of Illinois Press, Urbana, Illinoius.

Huelsenbeck, J. P., and Lander, K. M. 2003. Frequent inconsistency of parsimony under a simple model of cladogenesis. Syst. Biol. 52: 641--648.

Jenner, R. A. 2004. The scientific status of metazoan cladistics: why current reserach practice must change. Zoologica Scripta 33: 293--310.

Jiang, T. L., and Lawler, E. L. 1994. Aligning sequences via an evolutionary tree: computational complexity and approximation. In Proc. 26th ACM Symposium on the Theory of Computing, pp. 760--769. ACM, New York, New York.

Kjer, K. M. 1995. Use of rRNA secondary structure in phylogenetic studies to identify homologous positions: an example of alignment and data presentation from the frogs. Mol. Phylogenet. Evol. 4: 314--330.

Kluge, A. G. 1989. A concern for evidence and a phylogenetic hypothesis of relationships among Epicrates (Boideae, Serpentes). Syst. Zool. 38: 7--25.

Kluge, A. G. 1997a. Testability and the refutation and corroboration of cladistic hypotheses. Cladistics 13: 81--96.

Kluge, A. G., and Farris, J. S. 1969. Quantitative phyletics and the evolution of Anurans. Syst. Zool. 18: 1--32.

Kruskal, J. 1983. An overview of sequence comparison. In Time warps, string edits, and macromolecules. The theory and practice of sequence comparison (eds. D. Sankoff, and J.

Kruskal), pp. 1--44. CSLI Publications, Stanford, California (1999 reprint).

Kunin, V., and Ouzounis, C. A. 2003. The balance of driving forces during genome evolution in prokaryotes. Genome Research 13: 1589--1594.

Lee, D.-C., and Bryant, H. N. 1999. A reconsideration of the coding of inapplicable characters: assumptions and problems. Cladistics 15: 373--378.

Lutzoni, F., Wagner, P., Reeb, V., and Zoller, S. 2000. Integrating ambiguously aligned regions of DNA sequences in phylogenetic analyses without violating positional homology. Syst. Biol. 49: 628--651.

Maddison, W. P. 1993. Missing data versus missing characters in phylogenetic analysis. Syst. Biol. 42: 576--581.

Miklós, I., Lunter, G. A., and Holmes, I. 2004. A ``long indel" model for evolutionary sequence alignment. Mol. Biol. Evol. 21: 529--540.

Moilanen, A. 1999. Searching for most parsimonious trees with simulated evolutionary optimization. Cladistics 15: 39--50.

Moilanen, A. 2001. Simulated evolutionary optimization and local search: introduction and application to tree search. Cladistics 17: S12--S25.

Morgenstern, B. 2004. DIALIGN: multiple DNA and protein sequence alignment at BiBiServ. Nucl. Ac. Res. 32: W33--W36.

Morgenstern, B., Dress, A., and Werner, T. 1996. Multiple DNA and protein sequence alignment based on segment-to-segment comparison. Proc. Natl. Ac. Sc. 93: 12098--12103.

Murata, M., Richardson, J., and Sussman, J. 1985. Simultaneous comparison of three protein sequences. Proc. Natl. Ac. Sc. 82: 3073--3077.

Needleman, S. B., and Wunsch, C. D. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. J. Mol. Biol. 48: 443--453.

Nixon, K. C. 1999. The parsimony ratchet, a new method for rapid parsimony analysis. Cladistics 15: 407--414.

Nixon, K. C., and Carpenter, J. M. 1993. On outgroups. Cladistics 9: 413--426.

Nixon, K. C., and Little, D. P. 2004. The use of optimality criteria in DNA sequence data and its application in a new computer program (Abstract, 22nd annual meeting of the Willi Hennig Society, New York 2003). Cladistics 20: 90-91.

Notredame, C. 2002. Recent progress in multiple sequence alignment: a survey. Pharmacogenomics 3: 1--14.

Notredame, C., Higgins, D. G., and Heringa, J. 2000. T-COFFEE: A novel method for fast and accurate multiple sequence alignment. J. Mol. Biol. 302: 205--217.

Ochoterena, H. 2004. Independence of alignment and phylogenetic reconstruction and their optimality criteria (Abstract, 22nd annual meeting of the Willi Hennig Society, New York 2003). Cladistics 20: 91.

Phillips, A., Janies, D., and Wheeler, W. 2000. Multiple sequence alignment in phylogenetic analysis. Mol. Phylogenet. Evol. 16: 317--330.

Platnick, N. I. 1979. Philosophy and the transformation of cladistics. Syst. Zool. 28: 537--546.

Platnick, N. I., Griswold, C. E., and Coddington, J. A. 1991. On missing entries in cladistic analysis. Cladistics 7: 337--343.

Pleijel, F. 1995. On character coding for phylogeny reconstruction. Cladistics 11: 309--315.

Remane, A. 1952. Die Grundlagen des natürlichen Systems, der vergleichenden Anatomie und der Phylogenetik. Akademische Verlagsgesellschaft Geest & Portig, Leipzig.

Rieppel, O. C. 1988. Fundamentals of comparative biology. Birkhäuser Verlag, Basel.

Rieppel, O. 2003. Semaphoronts, cladograms and the roots of total evidence. Biol. J. Linn. Soc. 80: 167--186.

Rieppel, O., and Kearney, M. 2002. Similarity. Biol. J. Linn. Soc. 75: 59--82.

Rutishauser, R., and Sattler, R. 1989. Complementary and heuristic value of contrasting models in structural biology. III. Case study on shoot-like "leaves" and leaf-like "shoots" in Utricularia macrorhiza and Utricularia purpurea (Lentibulariaceae). Botanische Jahrbücher für Systematik 111: 121--137.

Saitou, N., and Nei, M. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. 4: 406--425.

Sankoff, D. 1975. Minimal mutation trees of sequences. SIAM J. Appl. Math. 28: 35--42.

Sankoff, D., and Cedergren, R. J. 1983. Simultaneous comparison of three or more sequences related by a tree. In Time warps, string edits, and macromolecules. The theory and practice of sequence comparison (eds. D. Sankoff, and J. Kruskal), pp. 253--263. CSLI Publications, Stanford, California (1999 reprint).

Sankoff, D., and Nadeau, J. H. (eds.) . 2000. Comparative genomics. Empirical and analytical approaches to gene order dynamics, map alignment and the evolution of gene families. Kluwer Academic Publishers, Dordrecht.

Sankoff, D., and Rousseau, P. 1975. Locating the vertices of a Steiner tree in an arbitrary metric space. Mathematical Programming 9: 240--246.

Sankoff, D., Cedergren, R. J., and Lapalme, G. 1976. Frequency of insertion-deletion, transversion, and transition in the evolution of 5S ribosomal RNA. J. Mol. Evol. 7: 133--149.

Sankoff, D., Morel, C., and Cedergren, R. J. 1973. Evolution of 5S RNA and the non-randomness of base replacement. Nature (New Biology) 245: 232-234.

Schwikowski, B., and Vingron, M. 1997. The deferred path heuristic for the generalized tree alignment problem. J. Comput. Biol. 4: 415--431.

Schwikowski, B., and Vingron, M. 2003. Weighted sequence graphs: boosting iterated dynamic programming using locally suboptimal solutions. Discr. Appl. Math. 127: 95--117.

Seitz, V., Ortiz García, S., and Liston, A. 2000. Alternative coding strategies and the inapplicable data coding problem. Taxon 49: 47--54.

Sellers, P. H. 1974. An algorithm for the distance between two sequences. J. Comb. Theory 16: 253--258.

Simmons, M. P. 2004. Independence of alignment and tree search. Mol. Phylogenet. Evol. 31: 874--879.

Simmons, M. P., and Ochoterena, H. 2000. Gaps as characters in sequence-based phylogenetic analyses. Syst. Biol. 49: 369--381.

Smith, T. F., Waterman, M. S., and Fitch, W. M. 1981. Comparative biosequence metrics. J. Mol. Evol. 18: 38--46.

Sokal, R. R. 1986. Phenetic taxonomy: theory and methods. Ann. Rev. Ecol. Syst. 17: 423--442.

Steel, M., and Penny, D. 2000. Parsimony, likelihood, and the role of models in molecular phylogenetics. Mol. Biol. Evol. 17: 839--850.

Strong, E., and Lipscomb, D. 1999. Character coding and inapplicable data. Cladistics 15: 363--371.

Thompson, J. D., Higgins, D. G., and Gibson, T. J. 1994. CLUSTAL W: improving the sensitivity of progresssive multiple alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucl. Ac. Res. 22: 4673--4680.

Thorne, J. L., Kishino, H., and Felsenstein, J. 1991. An evolutionary model for maximum likelihood alignment of DNA sequences. J. Mol. Evol. 33: 114--124.

Thorne, J. L., Kishino, H., and Felsenstein, J. 1992. Inching toward reality: an improved likelihood model of sequence evolution. J. Mol. Evol. 34: 3-16.

Tuffley, C., and Steel, M. 1997. Links between maximum likelihood and maximum parsimony under a simple model of site substitution. Bull. Math. Biol. 59: 581--607.

Vander Stappen, J., De Laet, J., Gama-López, S., Van Campenhout, S., and Volckaert, G. 2002. Phylogenetic analysis of Stylosanthes (Fabaceae) based on the internal transcribed spacer region (ITS) of nuclear ribosomal DNA. Plant Syst. Evol. 234: 27--51.

Vingron, M. 1999. Sequence alignment and phylogeny construction. In Mathematical support for molecular biology (eds. M. Farach-Colton, F. S. Roberts, M. Vingron, and M. Waterman), pp. 53--64. DIMACS series in discrete mathematics and theoretical computer science, Vol. 47.

Wang, L., and Jiang, T. 1994. On the complexity of multiple sequence alignment. J. Comput. Biol. 1: 337--348.

Wang, L., Jiang, T., and Lawler, L. 1996. Approximation algorithms for tree alignment with a given phylogeny. Algorithmica 16: 302--315.

Wheeler, W. 1996. Optimization alignment: the end of multiple sequence alignment in phylogenetics? Cladistics 12: 1--9.

Wheeler, W. 1998. Alignment characters, dynamic programming and heuristic solutions. In Molecular approaches to ecology and evolution (eds. R. DeSalle, and B. Schierwater), pp. 243--251. Birkhäuser Verlag, Basel.

Wheeler, W. C. 1999. Fixed character states and the optimization of molecular sequence data. Cladistics 15: 379--385.

Wheeler, W. 2001b. Homology and DNA sequence data. In The character concept in evolutionary biology (ed. G. P. Wagner), pp. 303--317. Academic Press, San Diego, California.

Wheeler, W. 2002. Optimization alignment: down, up, error, and improvements. In Techniques in molecular systematics and evolution (eds. R. DeSalle, G. Giribet, and W. Wheeler), pp. 55--69. Birkhäuser Verlag, Basel.

Wheeler, W. C. 2003a. Implied alignment: a synapomorphy-based multiple-sequence alignment method and its use in cladogram search. Cladistics 19: 261--268.

Wheeler, W. C. 2003b. Iterative pass optimization of sequence data. Cladistics 19: 254--260.

Wheeler, W. C., Gladstein, D., and De Laet, J. 2003. POY. Phylogeny reconstruction via optimization of DNA and other data. Version 3.0.11. Software and documentation freely available at ftp://ftp.amnh.org/pub/molecular/poy.

Wilkinson, M. 1995. A comparison of two methods of character construction. Cladistics 11: 297--308.

Yeates, D. 1992. Why remove autapomorphies? Cladistics 8: 387--389.