
Contents

Contributors	ix
1 Parsimony and phylogenetics in the genomic age <i>Victor A. Albert</i>	1
I Philosophical aspects of parsimony analysis, including comparison with model-based approaches	
2 What is the rationale for 'Ockham's razor' (a.k.a. parsimony) in phylogenetic inference? <i>Arnold G. Kluge</i>	15
3 Parsimony and its presuppositions <i>Elliott Sober</i>	43
II Parsimony, character analysis, and optimization of sequence characters	
4 The logic of the data matrix in phylogenetic analysis <i>Brent D. Mishler</i>	57
5 Alignment, dynamic homology, and optimization <i>Ward C. Wheeler</i>	71
6 Parsimony and the problem of inapplicables in sequence data <i>Jan E. De Laet</i>	81
III Computational limits of parsimony analysis: from historical aspects to competition with fast model-based approaches	
7 The limits of conventional cladistic analysis <i>Jerrold I. Davis, Kevin C. Nixon, and Damon P. Little</i>	119
8 Parsimony and Bayesian phylogenetics <i>Pablo A. Goloboff and Diego Pol</i>	148
IV Mathematical attributes of parsimony	
9 Maximum parsimony and the phylogenetic information in multistate characters <i>Mike Steel and David Penny</i>	163

V Parsimony and genomics

10 Using phylogeny to understand genomic evolution	181
<i>David A. Liberles</i>	
11 Dollo parsimony and the reconstruction of genome evolution	190
<i>Igor B. Rogozin, Yuri I. Wolf, Vladimir N. Babenko, and Eugene V. Koonin</i>	
References	201
Index	218

Parsimony and phylogenetics in the genomic age

Victor A. Albert

1.1 Parsimony inference

Parsimony (Ockham's razor) as a method of inference has a long history. Based upon a parsimony argument, Copernicus maintained that his heliocentric solar system theory was superior to the geocentric one of Ptolemy because of its greater simplicity. His reasoning was that Ptolemy's theory required what amounted to independent models for each planet's movement (extra parameters), whereas his own included the simplifying factor of Earth-Sun movement for each planet.¹

According to Copernicus, his theory "follow[s] Nature, who producing nothing vain or superfluous often prefers to endow one cause with many effects...." An important point about Copernicus's argument is that it represented an appeal to a universal law in nature, in other words, God. Modern considerations of parsimony methodology, especially those following Lamarck and Darwin, have by necessity been occupied with other, non-deist justifications (Sober 2003).

Parsimony today stands as a method of inference from observations. For example, if one has a coin with heads and tails, in the absence of any prior information about the coin other than this observation, the most parsimonious assumption for the result of a coin toss is one or the other, i.e. 50/50 chance. If the toss were to be repeated 1000 times, one could establish a frequency-based probability (with margin of error) that this were so.

If one were a Bayesian, and Joe had already tossed the coin 1000 times and gotten heads for 500 tosses, this prior probability could be used to assess the posterior probability.

In this simple example, parsimony, maximum likelihood (the mean of a normal distribution, as with 1000 coin tosses), and posterior probability all give the same answer. However, this was a *very* simple example, involving a single object with only two alternatives. The relationships between parsimony, likelihood, and Bayesian inference become much less obvious with more objects (characters) and alternatives (states). A biological example that Sober and Steel (2000) and Sober (2003) have examined is Crick's (1968) parsimony-based claim that all life has a common ancestor. Crick's argument was that since many different versions of the genetic code could have been possible, the common use of one (albeit with slight modifications) by all extant organisms strongly suggests their common ancestry. The idea is that selection would operate against code changes in descendants of a given code. In other words, one beginning of life with this attribute is more parsimonious than many (say, X). But Crick's is also a likelihood argument (Sober 2003):

$$\Pr(\text{code universal in extant organisms} | \text{one ancestor}) > \Pr(\text{code universal in extant organisms} | X \text{ separate ancestors})$$

which takes the standard form $\Pr(O|M_1) > \Pr(O|M_2)$, comparing likelihoods of O observations given models M . The formulation above follows the Law of Likelihood (Royall 1997), which states that a hypothesis with higher likelihood is preferable over one with lesser.

¹ Sober (e.g., Sober 1989, 2003; Chapter 3) has been an active student of this history, and I acknowledge his work for this example and several others I present below.

As pointed out by Sober (2003), parsimony and likelihood therefore provide identical evaluations of Crick's common ancestry hypothesis. However, further equivalence postulates between parsimony and likelihood, which I explore later, show the issue to be much more complex.

1.2 Examples of modern uses of parsimony

1.2.1 Curve fitting

Parsimony often plays a role in choosing among models fit to a set of points on an x,y plane. For example, a set of points might be regressed by a line a bit sloppily, or by a parabola far better. The question is, which model to accept? The latter requires an extra adjustable parameter, so it could be considered less parsimonious based on *economy of assumptions* (P_{EA} parsimony), i.e. Ockham's razor. On the other hand, a better *fit* to a parabola, which takes the form of minimization of residual variance between observations and model, is of course the likelihood, $\Pr(O|M)$.

There are different criteria to choose among models. In one example, an excellent parabolic fit (or even a higher-order one) might have no logical relationship to the data at hand (e.g. length of a naked mRNA strand vs. number of free bases after chemical degradation *in vitro*), and so a sloppy line would be the better model (through P_{EA}), albeit representing data that may have been collected in an error-prone manner.

Other data with better parabolic fit might have a realistic basis—this would then suggest a defiance of P_{EA} in terms of the number of adjustable parameters. But how to decide between the models? Two well-known criteria offer likelihood-based methods for this choice, the Bayesian and Akaike information criteria (BIC and AIC, respectively; see Sober 2003; Felsenstein 2004). In these criteria, parsimony takes the role of a penalty for complexity (in terms of number of adjustable parameters, p , referring to P_{EA}) among models, M , with different log-likelihoods. For example, with the BIC (the ratio of the average likelihoods for two models), for the most likely parabola to be preferred, it must fit the observations, n , better

enough than the most likely linear model to avoid the complexity factor, which is dependent on sample size:

$$\text{BIC}_M = -2\log L_M + p_M \log n$$

Roughly equal likelihoods, L , will likely mean that the line will win.

1.2.2 Trees of species or genes

Data points based on characters (e.g. nucleotides) sampled from species or genes can be analyzed under a *hierarchical model* in order to reconstruct most parsimonious trees. This operation is, of course, the central subject of this book. Most parsimonious trees are hierarchies or partially collapsed hierarchies with changes minimized across all characters that could show evidence for grouping. Here, groups are defined as two or more species or genes partitioned from two or more other species or genes. Not all character-state distributions can show evidence for grouping, and the specifics of information use is a major difference between parsimony, likelihood, and distance matrix methods. An illustration of this, as well as what trees demonstrate among the methods, will be useful.

For four species or genes, A, B, C, and D, there are 2^{n-1} different ways (in this case $n=4$) for binary characters to partition species or genes:

$$\begin{array}{ccccccc} \{AB\} & \{CD\} & \{ABC\} & \{D\} & \{ABCD\} & \{\} & \\ \{AC\} & \{BD\} & \{ABD\} & \{C\} & & & \\ \{AD\} & \{BC\} & \{ACD\} & \{B\} & & & \\ & & \{BCD\} & \{A\} & & & \end{array}$$

Parsimony can use only $2^{n-1} - (n+1)$ partitions, i.e. three—the two-item splits shown to the left. A character (in isolation from other characters) that argues for such a partition incurs one state change between such splits, yielding two groups (Fig. 1.1). None of the other partitions produce groups, although the middle four 3:1 splits incur state changes (these merely show a *difference*, between, say D vs. A, B, and C; Fig. 1.1). No changes are implied in the 4:0 split. However, likelihood methods use all of the eight partitions (see below). Distance matrix methods use information at rate $(n^2 - n)/2$. As species/gene number increases, it

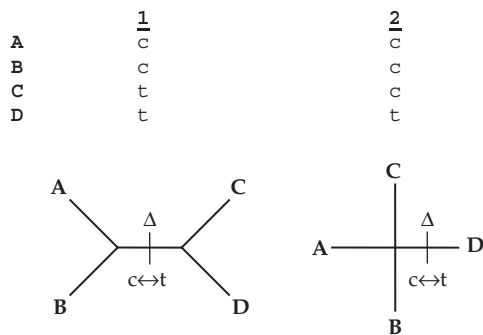


Figure 1.1 Examples of two characters and their states that (1) can show evidence for hierarchy vs. (2) evidence for difference. Trees implied by 1 vs. 2 alone are also shown. Δ indicates where a character-state change could occur. Note that only character 1 could support two groups, a group defined as comprising two or more items; character 2 can only support a fully collapsed tree in which one branch (not a group) is different from the others. A–D, species or genes. c/t, different pyrimidine bases.

can be seen that $2^{n-1} - (n+1)$ approaches 2^{n-1} , but that $(n^2 - n)/2$ lags far behind. Thus, for a given large n (such as with genomic-scale estimation of gene family phylogenies), parsimony utilizes the majority of all available evidence while only incorporating characters that could show evidence for grouping.

Likelihood trees demonstrate relationships among species or genes that maximize $\Pr(O|M)$, where M is an evolutionary model. As such, likelihood methods need *all* of the observations O , including the non-grouping partitions, to maximize the likelihood of the data; anything less would compromise the calculation. Branches of likelihood trees have lengths in terms of character-state change probabilities. Parsimony trees display relationships in terms of character-state changes along branches. Distance matrix methods of building trees show raw or model-adjusted differences between species or genes.

These illustrations are not meant as justifications for one method over the other in phylogeny reconstruction; rather, my goal has been to draw attention to differences in information use among the different methods, and to what trees derived from them demonstrate. This has often been confused in the theoretical and biological literature.

1.2.3 Phylogenetic models for which parsimony and likelihood are equivalent

I have already illustrated simple, non-phylogenetic models for which parsimony and likelihood are equivalent. The first attempts to establish this equivalence for phylogeny reconstruction were those of Farris (1973) and Felsenstein (1973). These models were different in that Farris's was basically Bayesian, with equal (flat) prior probabilities on all trees, whereas Felsenstein's was based on likelihood. Farris solved for the tree topology and character-state assignments at all points along branches, while including no assumption about rates of character-state change. Felsenstein's model summed over all possible character-state assignments, but required low rates of character-state change. According to Sober (Chapter 3), Farris's solution for topology *plus* character-state assignments (additional parameters) renders it inequivalent to likelihood, but that Felsenstein's parsimony model does achieve a likelihood equivalence. These considerations depend of course on the type of likelihood under consideration, for which there are several variants (Steel and Penny 2000; Goloboff 2003). To echo the point made by Steel and Penny (2000), both Farris's and Felsenstein's models are likelihood equivalents, just not for the same kind of likelihood.

Goldman's (1990) parsimony-likelihood formulation permits all branches to have the same length—a very simple model. However, Goldman, and indeed Sober (Chapter 3), assert its inequivalence with likelihood for basically the same reasons as for Farris's (1973) model: inference of the topology *plus something else*, in this case, ancestral character states. However, it is worth pointing out other views in the literature. According to Farris (1986) and Goloboff (2003a), ancestral states are not to be viewed as parameters:

Goldman (1990) decided that, even if the ancestral reconstructions are not parameters, they "could be treated as if they were." But they could also be treated (much more properly) as if they were not a parameter. The ancestral states are more like a kind of inferred observation (Farris, 1986). Parameters are instead those variables of the process that determine the conditions of the problem—the variables that determine the outcome of

evolution, that is. Even if not observed, the ancestral states are (just like observed states) part of that outcome. [Goloboff 2003a, p. 100]

Goldman's formulation of parsimony assumes that each character type occurs with a probability equal to the pathway with highest probability, among all the pathways that lead to that character type. If the probability of change in each branch is low, this estimation produces probabilities that are roughly proportional to the actual probabilities (i.e., the ones obtained by summing); that is, all the resulting character types are ranked in the same order of increasing probability by both criteria. This, however, does not convert the calculations under Goldman's model into estimations of a parameter; if a reconstruction was indeed a parameter, there would be one of them which would confer to the corresponding character type its true probability of occurrence under the model, and there is none. Only the sum of all reconstructions provides the true value for a given type. [Goloboff 2003a, p. 100]

Thus, Goloboff argues, using *the most likely reconstruction* (instead of the sum of likelihoods for all reconstructions) produces a good approximation of the actual likelihood, which is not exact, but then again some likelihood methods are not exact either. Goloboff gives the example that the assumption of nucleotide state frequencies remaining constant over time also implies that likelihood calculations are approximate instead of exact, since reconstructions are then not truly independent (they must sum to the assumed frequencies) (Goloboff 2003a, p. 101).

Without debate as an equivalence between parsimony and likelihood (Sober, Chapter 3; Goloboff 2003a) is the formulation of Tuffley and Steel (1997). They provided a proof that parsimony was a maximum likelihood estimator under the assumption of no common mechanism (potentially unequal change probabilities) for each character with r states and a symmetric change assumption. With this formulation, the different rates can either be very, very small or very, very large: in fact, only 0 or infinity, and nothing in between. The Tuffley and Steel result has been considered by some a *complex* parsimony equivalent because of its numerous adjustable parameters (the lengths of each branch for each character, as estimated from a single datum). On the other hand, Goldman's formulation is an extremely

simple model: the fit of a tree to data is solely based on its topology and on state change/stasis probabilities. As such, *parsimony inference can receive a likelihood equivalence at both ends of the complexity spectrum, which has been interpreted to speak toward its generality as an inferential method* (see Goloboff 2003a). Of course, equivalence between parsimony and likelihood between a few models does not mean that equivalence extends to *all* models, or that it has to do so in order to justify use of parsimony methods.

1.2.4 A non-likelihood justification for parsimony

Not all users of parsimony analysis care about equivalencies between parsimony and likelihood under certain process models. Farris himself, who produced a series of statistical interpretations of parsimony (1973, 1977, 1978), later downplayed these in one of the most important philosophical papers on parsimony analysis (Farris 1983). He famously stated that:

A number of authors, myself among them (Farris, 1973, 1977, 1978), have used statistical arguments to defend parsimony, using, of course, different models from Felsenstein's [1973]. . . . my own models, if perhaps not quite so fantastic as Felsenstein's, are nonetheless like the latter in comprising uncorroborated (and no doubt false) claims on evolution. If reasoning from unsubstantiated suppositions cannot legitimately question parsimony, then neither can it properly bolster that criterion. The statistical approach to phylogenetic inference was wrong from the start, for it rests on the idea that to study phylogeny at all, one must first know in great detail how evolution has proceeded. That cannot very well be the way in which scientific knowledge is obtained. [Farris 1983, p. 17]

Farris argued in favor of parsimony as a method that maximizes explanatory power among observations that could be expected to reflect genealogical relationships, i.e. potential homologies. He characterized most-parsimonious trees as the least falsified hierarchical hypotheses in the context of the philosopher Karl Popper's ideas on the treatment of observations (see Kluge, Chapter 2). However, Farris carried his argument into more general terms: *trees with minimal homoplasy*

(i.e. with minimal parallelism or reversal) must be preferred over trees that have more, because the latter require more observations to be dismissed for the sole purpose of protecting conclusions from 'offending' evidence. This is a fundamentally different use of the parsimony criterion, that which the data requires (P_{DR} parsimony), and indeed, this is a criterion that can be interpreted quantitatively. Operationally, for a given set of independent pairwise similarity statements,

$$\min \sum_k H_k \iff \max \sum_k J_k$$

where H and J represent independent statements of pairwise homoplasy and homology, respectively, across k characters (see De Laet, Chapter 6). Minimization of H , with the addition of a tree-independent constant, is equivalent to minimizing total steps (character-state changes) as calculated by standard parsimony software.

With reference to the minimization of H , Farris also refuted the commonly held belief that parsimony assumes rarity of homoplasy by use of an analogy to linear regression analysis; although residual variation in a least squares fit is certainly minimized, there is no requirement that this variation be small. Likewise, minimization of H occurs in the context of all characters, and this involves no requirement that estimated homoplasy be rare. For further discussion of Farris's arguments, see De Laet (Chapter 6; including an interesting elaboration) and Kluge (Chapter 2).

1.2.5 Parsimony and statistical consistency

Although consistency enters into further discussion below, I will only briefly deal with its formalities. As Felsenstein (2004; p. 107) explains:

An estimator is *consistent* if, as the amount of data gets larger and larger (approaching infinity), the estimator converges to the true value of the parameter with probability 1. If it converges to something else, we must suspect the method of trying to push us toward some untrue conclusion. In 1978 I presented...an argument that parsimony is, under some circumstances, an **inconsistent** estimator of the tree topology. [italics in the original; bold emphasis is mine]

Farris (1983) rejected Felsenstein's 1978 model by arguing against its applicability to real data. He didn't object to the general idea of seeking a consistent estimator; he just felt that one was not available in practice. A decade ago, colleagues and I modeled consistency for sequence evolution and concluded that the 'Felsenstein zone' of inconsistency (under Felsenstein's own conditions) was small enough to be insignificant for real data (Albert *et al.* 1992, 1993). Felsenstein showed similar results himself (2004; see also Steel and Penny 2000), which echo our findings that as r states increase, the zone of inconsistency decreases. This will be seen to have bearing when gene-order data are discussed below.

1.2.6 Other practical considerations.

Parsimony analysis yields hierarchic results that are both fully diagnosable and interconvertible with the original data (Fig. 1.2). This is a very positive feature in terms of tree interpretation and for information storage and retrieval (Farris 1979). As stated above, most-parsimonious trees have branch lengths in terms of changes among the states of informative characters. This format is more intuitive than branches in terms of state-change probabilities, distances (via some metric), or those with no dimensions whatsoever. Indeed, many investigators have exploited parsimony's branch-length properties to optimize their original data on to trees derived from other methods; however, such comparisons are not interconvertible with the data matrix, rendering interpretations potentially *tree*-biased as opposed to *data*-biased (remember that the *data* that go into most-parsimonious trees provide the branch lengths that come back out; see Mishler, Chapter 4).

1.3 Genomic-scale data and parsimony

Current whole-genome sequences and projects underway represent the tip of the iceberg. Now that we have complete genome sequences for many prokaryotes, several eukaryotes, and numerous organellar DNAs, bioinformaticians

```

O cc c c c c c c cccccccccc c c c c c c cccccccccc
A gg g g g c c c gggggggggg c c c c c c cccccccccc
B gg g g g c c c cccccccccc g c c c c c cccccccccc
C gg g g c c c c cccccccccc c g c c c c cccccccccc
D gg g c c c c c cccccccccc c c g c c c cccccccccc
E gg c c c g c c cccccccccc c c c g c c cccccccccc
F gg c c c g g c cccccccccc c c c c g c cccccccccc
G gg c c c g g g cccccccccc c c c c c g cccccccccc
H gg c c c g g g cccccccccc c c c c c gggggggggggg
  1  2  3  4  5  6  7

```

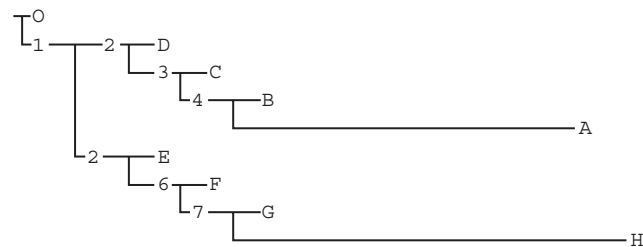


Figure 1.2 Phylogenetic trees based on parsimony are fully diagnosable. Similarities that had the capacity to bear hierarchical information, here characters 2–7, were those used to build the most parsimonious tree, and upon inspection it can be seen that inferred character-state changes can be readily optimized on to internal branches. All other characters shown in this example only show difference, as opposed to evidence for hierarchy. This fact can be readily appreciated by examination of the lengths of branches—two differences separate O (outgroup) from the other species or genes A–H; A and H have whole blocks of singular differences, B–G each have singular differences assigned, and all of this is reflected in the tree as differential branch lengths, not as arguments for a different overall hierarchy. Note also that the tree and its branch lengths are fully interconvertible with the original data matrix.

must find ways to reduce this complexity to provide meaningful fodder for biological hypotheses. One urgent need is for fast and predictive phylogenetic estimation of species and gene relationships. Parsimony is a method that has the logical and practical attributes discussed above, as well as, recently, the speed necessary to carry out massive topological calculations.

1.3.1 Sequence data and tree size

Parsimony analyses of sequence data for large numbers of species have been possible for a number of years (see Davis *et al.*, Chapter 7), but only recently have these become fast enough to be considered of use at the genomic scale. For genome comparisons, it will be important to use parsimony calculations to determine gene relationships within gene families or superfamilies. The issue of orthology vs. paralogy is important in the context of molecular evolutionary hypothesis testing (see Liberles, Chapter 10, and Rogozin *et al.*, Chapter 11). One current application, TNT (Goloboff *et al.* 2004), has the ability

to solve reliably for large most-parsimonious trees that were once thought to be intractable problems. For example, the 500-sequence *rbcl* data set for seed plants (see Davis *et al.*, Chapter 7) can now be solved for most-parsimonious trees in seconds (<42 s on my 1.6 GHz Pentium M laptop, in fact). This application has also been parallelized (Goloboff *et al.* 2003b), so for individual sequence alignments, the practical limits for genomic-scale sequence data will be the strength of such alignments.

In the case of *rbcl*, a highly conserved protein-coding gene with no introns, sequence start-stop and internal base alignment is unambiguous. However, this is certainly not a generalization that can be made for genes in general, not to mention (e.g.) non-coding regions in between genes. This begs an issue that I have avoided until now—according to Wheeler (Chapter 5) and De Laet (Chapter 6), the logic of *a priori* multiple alignment is erroneous and the results are incomplete at best. These authors argue persuasively for *optimization of sequences as whole, complex characters* under Sankoff parsimony (Sankoff 1975). However,

practical limits rise greatly for such algorithms, which create compound NP-complete problems that must, by current necessity, be solved heuristically.

Likelihood- or distance-based phylogenetics will reach different sorts of complexity blockades in dealing with relationships among large gene families. For example, the R2R3-MYB gene family in *Arabidopsis* is composed of ca. 100 members, as opposed to only three found in human (Martin and Paz-Ares 1997; Romero *et al.* 1998). Imagine ca. 100 R2R3-MYBs across 500 plant genomes—perhaps 50 000 genes, which given advances in sequencing technology, isn't a far-fetched possibility in the not-too-distant future. If full diagnosability (tree/matrix interconvertability) and speed of execution together form the important criterion, distance matrix and likelihood methods will prove inferior. Distance matrix methods are fast, but they decompose information into pairwise estimates of path lengths between n_i and n_j and then, at best, try to reassemble them into a tree that somehow optimizes these lengths. Such an operation is fraught with error since pairwise path lengths may show no relationship to those on reconstructed trees. Moreover, the inherent information loss, especially as n gets large, is unacceptable (see above). Likelihood methods do not provide character-state diagnoses either, and are understandably slower than parsimony given that calculations are CPU-intensive, especially as n increases (even using the pruning algorithm, computational effort is proportional to $k(n-1)r^2$; Felsenstein 2004). Supercomputers and CPU clusters speed up likelihood calculations, but eventually a tradeoff will be reached. Besides, to quote Felsenstein (2004, p. 122):

If it escapes the clutches of long branch attraction [inconsistency], parsimony is a fairly well-behaved method. It is close to being a likelihood method, but is simpler and faster. It is robust against violations of the assumption that rates of change at different sites are equal. (It shares this with its likelihood doppelgänger.)

Given the equatability of parsimony with likelihood under several models that range from simple to complex (see above), parsimony should be the method of choice as applied to genomic-scale questions that include enormous numbers of

species or genes. It will no doubt arise in some readers' minds, however, that the issue of tree size, e.g. for sequences of 50 000 genes, could impact statistical consistency. Indeed, some have cautioned that inconsistency can occur more often as trees become larger and larger (Kim 1996). I will present a more positive outlook on parsimony and large trees below.

1.3.2 A conjecture on parsimony and large phylogenetic trees

Background

With reference to the largest phylogenetic analysis yet attempted (2538 *rbcL* sequences for photosynthetic organisms), colleagues and I (Källersjö *et al.* 1999) observed that relatively rapidly evolving nucleotide sites, such as those in third positions of codons, provide the majority of tree structure despite initial estimates of saturation and high levels of homoplasy on most-parsimonious trees. We pointed out in reference to this analysis that, analyzed by themselves, third positions resolve 1327 supported groups with an average parsimony jackknife frequency of 85%, whereas the first two positions together resolve only 431 groups, with an average frequency of only 75%. The groups recovered by third positions are also well supported by the full data and are spread over the tree, including both older and younger lineages. In contrast, the first two positions fail, for example, to recognize either land plants or flowering plants as monophyletic groups.

We also generated random subsets (10 for each) of $n=100$ species, $n, 2n \dots 10n$, from the 2538-species matrix, and calculated the average retention index, position-wise within codons, for each subset. The retention index for individual characters— $(g-s)/(g-m)$, where g is the maximum number of steps, s is the most-parsimonious number, and m is the minimum number—measures the amount of initial similarity retained as homology on most parsimonious trees (Farris 1989a). As matrix size rose from 100 to 1000, the retention index rose for third positions as matrix size increased. In contrast, the retention indices for first and second positions—those sometimes

avored for molecular phylogenetics because they evolve more slowly—decreased. Our simple interpretation, also following the group support data reported above, was that third positions were performing better than first or second positions. Moreover, with respect to total homoplasy, the consistency index (m/s ; Kluge and Farris 1969), which is inversely proportional to the total number of substitutions, gave the converse view: first, second, and third positions averaged 0.155, 0.178, and 0.046, respectively on the tree calculated from all positions—that is, *greatest homoplasy was discovered for the more rapidly evolving third positions, despite their better performance*. Given these results, our conclusion was that homoplasy can *increase* phylogenetic structure.

Conjecture

Homoplasy on trees can have a direct relationship with rates of change. It certainly does on large trees, such as those discussed above ($n = 100\text{--}2538$), for which enough branches exist to observe the products of high evolutionary rates. The branches of small trees (e.g. $n = 4\text{--}6$), such as those often used for simulation studies, would not be expected to reveal underlying rate differences as accurately for vast divergences, and such differences are precisely those that might lead to inconsistency.

In Chapter 9 of this book, Steel and Penny prove a common-mechanism equivalence between parsimony and likelihood when the number of character states, r , is large enough (we will get back to this issue later regarding gene-order data). They also show that under such conditions, few characters, k , are required to arrive at a most-parsimonious solution.

My conjecture is that a parsimony-likelihood equivalence can hold when r is much smaller than required by Steel and Penny's Theorem 9.6.2, e.g. in the $r = 4$ case as for nucleotide data, if n is large enough.

Erdős *et al.* (1999) proved that a compatibility tree, from which homoplasy is prohibited, requires at least $(n - 3) * \log(n - 3) - (n - 3)$ informative characters (i.e., with grouping potential) to reconstruct a tree of n species or genes with at least 50/50 probability. For the 2538 species case illustrated

above, this quantity is at least 17300 informative sites. But, in reality, k was only 1428—the number of bases in the *rbcL* gene. Erdős *et al.* (1999) also established that at least some phylogenetic methods should require, for a given constant c , only $c \log(n)$ or at worst a power of $c \log(n)$ characters. Steel and Penny's Conjecture 9.4.3 and Proposition 9.4.4 (Chapter 9) similarly suggest that k follows logarithmic growth on n , and as n grows, so too would homoplasy for some characters, especially given any limitation put on r . This requirement for homoplasy echoes the empirical findings discussed above. With Mike Steel's help, I provide below a mathematical formalization of my conjecture in light of these findings.

Conjecture 1.3.1. Consider the r -state symmetric Poisson model. For any $\epsilon > 0$ and constant $B \geq 1$ there exist constants h and c that depend only on r , ϵ , and B for which the following holds.

Suppose k characters are generated independently for this model on any fully resolved phylogenetic tree T with n species or genes for which all the branch lengths of T are at most h and the ratio of any two branch lengths of T is at most B . Then provided $k \geq c \log(n)/f^2$, where f is the smallest branch length in T , maximum parsimony will correctly recover T with probability at least $1 - \epsilon$.

Here, ϵ is any real number, and $B = 1$ is the case where all branch lengths are equal. As f , which is a function of n , gets smaller and smaller, the sequence length needed to 'detect' that branch has to grow (indeed quadratically with $1/f$). The role of B is to avoid inconsistency (as described by Felsenstein 1978a). However, note that the conjecture just says 'for any B ', so one could take $B = 1000$ *a priori*, and thereby allow one branch to be 1000 times as long as another. This will impact the values of c and h , but these are constants so far as n is concerned. Presumably also as n increases, the value of B may come down closer to 1, provided that adding to n does not create a branch that is too short.

The conditions stated mean that most of the characters will have a fair degree of homoplasy—indeed, the expected number of steps will go to infinity with n , since each branch length is bounded below by h , which is a positive number.

1.3.3 Strings, and more on r -state characters

Around the same time, Steve Farris and I developed methods that were intended to lower worries about parsimony and inconsistency. I will begin with my procedure (Albert *et al.* 1994), which was to accept strings of nucleotides (randomly selected) as unit characters instead of individual bases; these strings were recoded as presence vs. absence for data analysis. The intended effect was to reduce the probability of homoplasy given, e.g. that a six-base pair string is much easier to lose once it exists than it is to regain once lost.

My argument was based on investigations of Dollo parsimony (Farris 1977) and its use with DNA restriction-site data (Albert *et al.* 1992). In this context, Dollo *never* permits parallel gains of a restriction site (often a six-base recognition string), only multiple losses. In this earlier work, we concluded that the Dollo model was too severe and that despite the asymmetry in probabilities just discussed, parsimony with equal character-state weights (Kluge and Farris 1969; Farris 1970) was more appropriate. However, my work did not consider increasing string length. Felsenstein (2004, p. 236) has examined the parallel gain case more thoroughly, solving for the probability (under the Jukes–Cantor model) that two species or genes and their common ancestor each have/had (+) a particular nucleotide string of k sites given substitution rate q and t units of branch length:

$$\Pr(+++) = \left(\frac{1}{4}\right)^k \left(\frac{1 + 3e^{-\frac{3}{4}qt}}{4}\right)^{2k}$$

The conditional probability that the ancestor is in state + is the ratio of this equation with the probability that both species/genes have state +. The latter probability takes the exact form as above while replacing $2qt$ for qt and the exponent k for $2k$. This probability ratio clearly demonstrates that as strings grow longer and longer, the probability of parallel gain still remains small so long as substitution rates remain low.

This is precisely why I developed the string character method; if one were to code *only* those completely matching strings beginning at certain nucleotide positions, especially larger and larger ones, then these should be rather conservative characters for deep branchings within phylogenetic problems. As such, the string character concept

need not be restricted to nucleotide data; amino acid data, exons/introns, or even genomic regions could be so coded (see below).

Now on to r -state characters. Farris and Källersjö presented a related method, supersites, at the 1999 meetings of the Willi Hennig Society in Göttingen, Germany. With supersites, strings of nucleotides are recognized beginning at nucleotide W and then parsed downwards through the matrix, recognizing as many character states as necessary to account for differences within the strings. Supersites can therefore generate considerable character-state space among fewer informative characters. However, Steel and Penny (2000) suggest that such procedures may not avoid inconsistency because probabilities of change along branches increase $\rightarrow 1$ as a function of k , where $r = d^k$ and d is the number of possible character states.

1.3.4 Gene content

Genomic-scale phylogenetic studies based on gene content are reviewed by Rogozin *et al.* (Chapter 11). Two approaches have been used: (1) estimate species trees from orthologous gene presence vs. absence among whole genomes, or (2) optimize these data on to a predetermined species tree. In my string character method, above, I permitted string presence-absence to have equal weight, whereas the probabilities modeled above imply asymmetric weights favoring losses. So which character-state weights to use? Rogozin *et al.* (Chapter 11) discuss Dollo analyses based on whole genes, which are of course nucleotide strings themselves (see above). The Dollo assumption is the asymptotic case, and with reference to the equation above this should be entirely appropriate as string size increases, say, to 1428 bases. Use of Dollo optimization onto species trees incorporates the same state-change asymmetry. Moreover, Huson and Steel (2004) have shown that Dollo parsimony compares very favorably with a genesis-loss likelihood model they constructed to analyze gene-content data.

1.3.5 Gene order

The growing rate of whole-genome sequencing, particularly for the relatively small and circular genomes (prokaryote, chloroplast, and most mitochondrial), has been accompanied by heightened interest in determining phylogenetic relationships

based on gene order (synteny). The problem is not a simple one, at least in terms of encoding the data. For one, Steel and Penny point out in Chapter 9 that the order of G genes in a signed (oriented) circular genome can display any of $2^G(G-1)!$ combinations. Nonetheless, their proof of equivalence between likelihood and parsimony for large character-state space states bodes well for use of computationally simpler parsimony calculations in this genomic arena.

Parsimony analyses of gene order are related to analyses of string data (as discussed above), but differ in their attempt to account for *adjacency vs. non-adjacency* of strings. Coding methods for use with parsimony analysis have already been under investigation, e.g. Maximum Parsimony on Multistate Encodings (MPME; Wang *et al.* 2002, as suggested by Bryant 2000) methods. This method produces signed, multistate circular permutations of gene adjacency on circular genomes (see discussion in Steel and Penny, Chapter 9). In one simulation study, MPME has been shown to have greater accuracy in comparison with a method incorporating neighbor joining (Wang *et al.* 2002), a distance matrix method that inherently incorporates less information from the data (see above). Still other coding methods exist or are under development, including a technique that utilizes Dollo parsimony on tightly linked gene pairs that are then binary-recorded (Wolf *et al.* 2001; see Rogozin *et al.* Chapter 11). This method is directly related to the string recoding method of Albert *et al.* (1994).

A limitation encountered by Wang *et al.* (2002) and mentioned by Steel and Penny (Chapter 9) is the relatively small number of character states permitted by the most rigorous parsimony software (e.g. TNT) and required by the multistate coding methods. In other words, getting anywhere with the gene-order issue will require algorithmic advances regarding state space.

1.3.6 Microarray data

Hierarchic analysis of microarray expression data has become routine. However, almost all methods used are those of phenetic clustering (Eisen *et al.* 1998), which only supplies dimensionless levels of difference based on a distance matrix of

log-transformed fluorescence intensities. Clusters of genes or 'treatments' (including tissue types) are formed based on shared patterns of up- vs. down-regulation of gene expression. Such analyses have proven of some use in (e.g.) tumor classification by gene-expression profiles, as well as, by inverting the matrix, identification of genes active in different tumor types. The analyses do not intend to be phylogenetic. However, phylogenetic methods, such as parsimony analysis, can be brought to bear on microarray data, at least when these data could be expected *a priori* to show evidence for hierarchy (e.g. through hereditary relationship). A parsimony approach to microarray analysis has been developed (Planet *et al.* 2001; Sarkar *et al.* 2002) and applied to the tumor classification problem. However, classification of different tumor types may not fit a phylogenetic model; gene regulation can be hierarchical and is certainly heritable, but it may also be networked, and tumor types do not share clear evolutionary relationships. A less stringent view on fit to model might be worth adopting for exploratory studies, since Sarkar *et al.* did identify gene-expression events that had also been identified by phenetic clustering.

There is one study of which I am aware that explicitly used parsimony analysis to reconstruct heritable relationships (they cited Planet *et al.* 2001). Uddin *et al.* (2004) used genome-wide expression profiles from primate brains to perform a parsimony analysis of organismic relationships; echoing other substantial evidence (e.g. Salem *et al.* 2003), the chimpanzee was identified as *Homo sapiens'* closest relative. Another classic study of heritable gene expression relationships within and between species was that of Oleksiak *et al.* (2002) on *Fundulus* fish populations. These authors used the phenetic clustering methods of Eisen *et al.* (1998) to group populations by genes, as well as genes by populations. Although Oleksiak *et al.* were studying population differentiation and not phylogeny *per se*, it would have been possible to use parsimony methods that incorporate population-level information. Sarkar *et al.*'s Characteristic Attribute Organization System (CAOS) is closely related to Population Aggregation Analysis (PAA; Davis and Nixon 1992), which identifies patterns of discrete features that

unambiguously mark groups (i.e. that have gone to fixation). The CAOS approach further identifies characteristic expression patterns found in some members of a group and never outside that group, those found in all members of a group but never found together outside that group, and those found in some samples within a group but never outside that group.

CAOS illustrates the major advantage of using parsimony analysis for microarray data: diagnosability (see above). State changes that identify groups (of genes or treatments) and changes among members of groups have far greater predictive value than dimensionless clustering.

1.4 The future: some predictions

It is difficult to predict future modes and rates of genomic-scale data acquisition, but with Moore's law, computer capacity should open up previously inaccessible data-analysis possibilities in less than a decade. Parsimony will remain an indispensable part of the phylogenetics and genomics tool kit, particularly to estimate enormously large trees with full diagnosability, and also for data with large character-state space (e.g. gene-order information). Perhaps a proof for Conjecture 1.3.1 can be given, establishing that particular amounts of homoplasy on large-enough trees

can render another parsimony-likelihood equivalence. A massive increase in whole-genome sequencing will no doubt permit refinements to estimations of ancestral gene content. To mention an area barely discussed in this chapter, optimization of whole sequences as complex characters will also become a practical and everyday tool with large numbers of species or genes. The gene-order issue, which will no doubt develop further with different encoding methods, will also include similar approaches to optimization of whole genomes as complex characters (see De Laet, Chapter 6). Finally, parsimony analyses of microarray data should become commonplace for gene-expression data with underlying hereditary relationships, such as for phylogenetic and population genomics.

1.5 Acknowledgments

I thank Pablo Goloboff for his extremely helpful comments on the manuscript, Mike Steel for his interest in helping me formalize Conjecture 1.3.1, and Steve Farris for the example in Fig. 1.2. Again, the work of Elliott Sober is acknowledged for useful examples. I also thank the authors of the various chapters of this book; their insights and topical reviews helped make this introductory chapter possible.

