J Mol Evol (1998) 47:686–690

JOURNAL OF MOLECULAR EVOLUTION

Problems with Parsimony in Sequences of Biased Base Composition

Adam Eyre-Walker

Centre for the Study of Evolution & School of Biological Sciences, University of Sussex, Brighton BN1 9QG, UK

Received: 13 November 1997 / Accepted: 18 June 1998

Abstract. Parsimony is commonly used to infer the direction of substitution and mutation. However, it is known that parsimony is biased when the base composition of the DNA sequence is skewed. Here I quantify this effect for several simple cases. The analysis demonstrates that parsimony can be misleading even when levels of sequence divergence are as low as 10%; parsimony incorrectly infers an excess of common to rare changes. Caution must therefore be excercised in the use of parsimony.

Key words: Parsimony — G + C content — Base composition — Substitution pattern — Mutation pattern

Introduction

Parsimony is one of the most commonly used procedures in evolutionary analysis. It has been used extensively in the reconstruction of phylogeny and is often used in a number of other analyses. Here I consider its use in the inference of substitution and mutation patterns.

Let us imagine that we wish to infer the pattern of substitution along a lineage. The easiest method is to get two or more outgroup sequences and use parsimony to infer the direction of substitution. For example, if the sequence under consideration has a T at a site, but two outgroup sequences a C, then we would infer that a C → T substitution had occurred. Similarly, we could use parsimony and outgroup sequences to infer the mutation by which a polymorphism arose; if there were C's and T's segregating at a site in a population, but an outgroup sequence had a C, we would infer that the polymorphism arose via a C → T mutation.

Unfortunately parsimony can be misleading in sequences of biased base composition (Collins et al. 1994; Perna and Kocher 1995). Parsimony tends to reconstruct ancestral states as one of the common nucleotides; parsimony therefore overestimates the number of common to rare changes. The purpose of this paper is to quantify the biases introduced by parsimony in the inference of substitution and mutation patterns and, hence, give some direction as to when parsimony can be used, and when caution should be excercised.

Collins et al. (1994) previously considered the problem of inferring the substitution pattern; they showed, using simulations, that parsimony infers an excess of common to rare changes, particularly if the level of divergence is large. However, they considered a rather complicated model with nine taxa, in which the level of divergence was not easily inferred. Here I consider some simple cases analytically. The cases are of the sort generally encountered in the inference of substitution and mutation patterns (e.g., Gojobori et al. 1982; Imanishi and Gojobori 1992; Tamura and Nei 1993; Ballard and Kreitman 1994; Akashi 1995). The analyses show that, even with modest levels of sequence divergence, parsimony can be misleading.

Basic Theory

For simplicity let us assume that there are just two states at each site, which we denote 1 and 2. Let 1 → 2 substitutions occur at a rate of α per generation and 2 → 1 substitutions at a rate of β. It is straightforward to show that the equilibrium frequency of state 1 is

Correspondence to: author; *e-mail:* a.c.eyre-walker@sussex.ac.uk

$$f = \frac{\beta}{(\alpha + \beta)} \tag{1}$$

and the rate of evolution

$$R = f\alpha + (1 - f)\,\beta = 2f(1 - f)k \tag{2}$$

where $k = \alpha + \beta$. *If $X_1[t]$ is the probability that a site which is in state 1 at time 0 is 1 at time t,* then the change in $X_1[t]$ per generation is

$$\Delta X_1[t] \;=\; X_1[t](1 - \alpha) + (1 - X_1[t])\beta - X_1[t] \tag{3}$$

This can be approximated by the continuous function

$$\frac{\partial X_1[t]}{\partial t} = X_1[t](1 - \alpha) + (1 - X_1[t])\beta - X_1[t] = \beta - X_1[t]k \tag{4}$$

Solving this, noting that $X_1[0] \;=\; 1$, and using Eq. (1), we get

$$X_1[t] \;=\; f + (1 - f)e^{-kt} \tag{5}$$

By similar reasoning it can be shown that

$$X_2[t] \;=\; (1 - f) + fe^{-kt} \tag{6}$$

(Eyre-Walker 1994)

Inferring the Pattern of Substitution

Let us consider the pattern of substitution inferred using parsimony in a lineage when we have three sequences. Let us assume that rates of evolution are the same along each lineage and that the sequences split simultaneously (i.e., they form a star phylogeny). Under parsimony we infer the 1 → 2 substitutions from two types of sites: sites which were ancestrally 1, but which are now in state 2 in one of the lineages, and sites which were in state 2 ancestrally, but are now in state 1 in two of the lineages. The probabilities of these are $3X_1[t]^2(1 - X_1[t])$ and $3(1 - X_2[t])^2\, X_2[t]$. Overall the number of substitutions inferred to be 1 → 2 and 2 → 1 is therefore

$$P_{12}(3) = 3(\,fX_1[t]^2\,(1 - X_1[t]) + (1 - f)(1 - X_2[t])^2\,X_2[t]) \tag{7}$$

$$P_{21}(3) = 3(f(1 - X_1[t])^2\,X_1[t] + (1 - f)X_2[t]^2(1 - X_2[t]))$$

The factor of three is lost from these equations if we consider the pattern of substitution along a specific lineage.

Since the sequences are stationary in composition, there are equal numbers of 1 → 2 and 2 → 1 substitutions. However, parsimony infers that there are many more 1 → 2 substitutions than 2 → 1 substitutions when there is a bias towards state 1 (Fig. 1a). The bias is surprisingly large: even with relatively low levels of divergence and modest amounts of compositional bias, parsimony infers that there are many more 1 → 2 substitutions than 2 → 1 substitutions; for example, if the frequency of state 1 is 0.8 and the divergence is 0.1 (i.e., 10%) along each lineage, then we would infer that 61% of the substitutions were 1 → 2 changes.

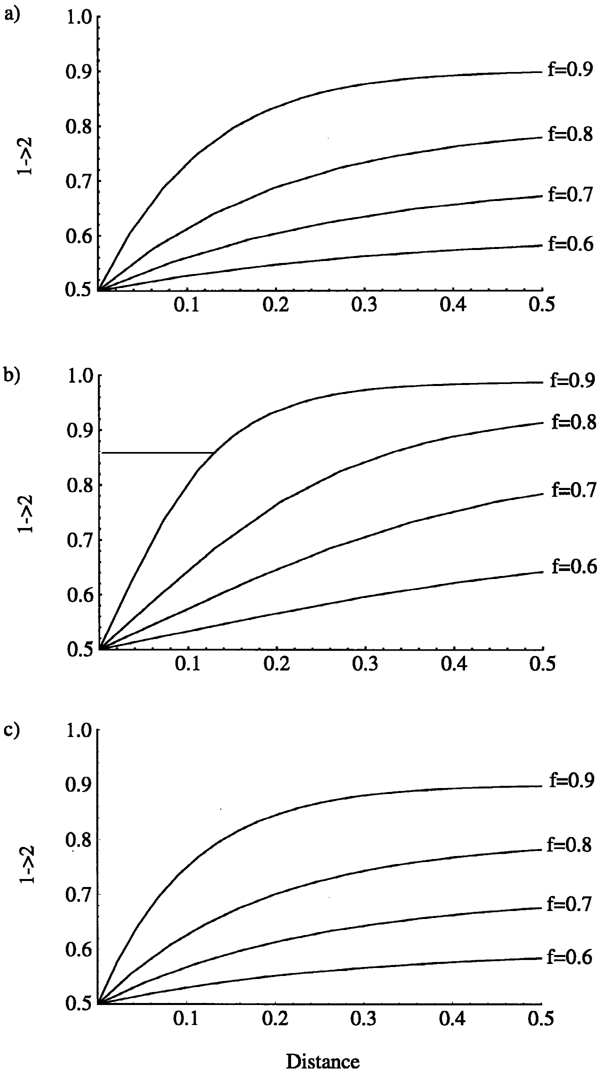

Fig. 1. The proportion of substitutions inferred by parsimony to be 1 → 2 changes when there are **(a)** three or **(b)** four sequences in a star phylogeny or **(c)** three sequences in a nonstar phylogeny where the ingroup sequences diverged at $t/2$ generations. The figures at the *right* give the frequency of state 1.

Surprisingly the situation does not improve with the addition of another sequence. Assuming a star phylogeny and equal rates of evolution along each lineage, the number of substitutions inferred for four sequences is

$$P_{12}(4) = 4(\,fX_1[t]^3(1 - X_1[t]) + (1 - f)(1 - X_2[t])^3\,X_2[t]) \tag{8}$$

$$P_{21}(4) = 4(f(1 - X_1[t])^3\,X_1[t] + (1 - f)X_2[t]^3(1 - X_2[t]))$$

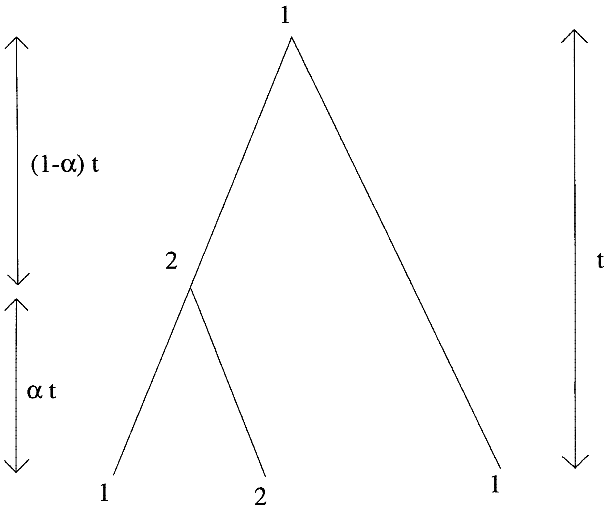

Fig. 2. Inferring the substitution pattern when there is a nonstar phylogeny. A 1 → 2 substitution inferred to have occurred within the last αt time units along one of the ingroup lineages.

The difference between the number of 1 → 2 substitutions inferred by parsimony and the number of 2 → 1 substitutions is even greater than with three sequences (Fig. 1b); for the example above (divergence of 10%, frequency of state 1 0.8), parsimony now infers that 64% of the substitutions were 1 → 2. The situation improves with five sequences; the biases are slightly lower than those obtained with three sequences (results not shown).

Generally we do not have a star phylogeny. Let us consider the problem of inferring the pattern of substitution when we have two ingroup sequences and an outgroup sequence. Imagine that the outgroup sequence diverged t, and the ingroup sequences αt generations in the past (where $0 < \alpha < 1$) (Fig. 2). There are four pathways which would lead us to infer that a 1 → 2 substitution had occurred in one of the ingroup lineages; one of these is shown in Fig. 2. The probability of this pathway is $X_1[t](1 - X_1[(1 - \alpha)t])(1 - X_2[\alpha t])X_2[\alpha t]$. The overall number of 1 → 2 and 2 → 1 substitutions inferred from the data is therefore

$$\begin{aligned} P_{12}(3)^* = {} & 2f\{X_1[(1-\alpha)t]X_1[\alpha t](1-X_1[\alpha t])X_1[t] \\ & + (1-X_1[(1-\alpha)t])(1-X_2[\alpha t])X_2[\alpha t]X_1[t]\} \\ & + 2(1-f)\{X_2[(1-\alpha)t]X_2[\alpha t](1-X_2[\alpha t]) \\ & (1-X_2[t]) + (1-X_2[(1-\alpha)t]) \\ & (1-X_1[\alpha t])X_1[\alpha t](1-X_2[t])\} \end{aligned} \tag{9}$$

$$\begin{aligned} P_{21}(3)^* = {} & 2f\{X_1[(1-\alpha)t]X_1[\alpha t](1-X_1[\alpha t])(1-X_1[t]) \\ & + (1-X_1[(1-\alpha)t])(1-X_2[\alpha t])X_2[\alpha t] \\ & (1-X_1[t])\} + 2(1-f)\{X_2[(1-\alpha)t]X_2[\alpha t] \\ & (1-X_2[\alpha t])X_2[t] + (1-X_2[(1-\alpha)t]) \\ & (1-X_1[\alpha t])X_1[\alpha t]X_2[t]\} \end{aligned}$$

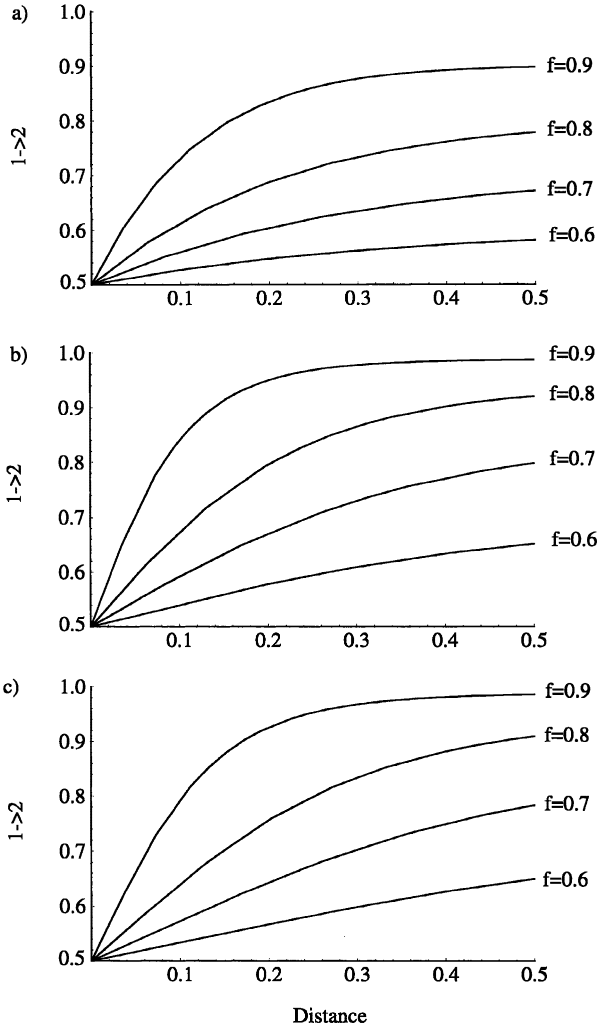

Fig. 3. The proportion of polymorphisms inferred by parsimony to have arisen by a 1 → 2 mutation using **(a)** one and **(b)** two outgroup sequences in a star phylogeny and **(c)** two outgroup sequences in a nonstar phylogeny where the more recent outgroup diverged $t/2$ generations ago. The figures at the *right* give the frequency of state 1.

In Fig. 1c the proportion of substitutions inferred to be 1 → 2 changes is plotted against the divergence that has occurred along the lineage since the root (at time 0) for the case where the ingroups diverged at $t/2$. It will be seen that the situation is almost-identical to the star phylogeny case; in fact the number of 1 → 2 changes that are inferred for the nonstar phylogeny is slightly greater for a given divergence. At first sight this appears counterintuitive; if the ingroups are closely related (i.e., small α), we expect to have more information about the ancestral state and to be able to infer the pattern of substitution more accurately. However, this added information is almost exactly compensated for by a loss of information from the outgroup sequence; note that if the time of divergence for the ingroup sequences is αt, the total time from the node joining the two ingroup sequences to the outgroup sequence is $t + (1 - \alpha)t$.

Inferring the Direction of Mutations

Parsimony has also been used, on occasion, to infer the direction of mutation by which a polymorphism arose (e.g., Imanishi and Gojobori 1992; Ballard and Kreitman 1994; Akashi 1995). Let us assume for simplicity that the mutations are neutral, then the rate at which 1 → 2 polymorphisms arise per site can be written as $\Lambda\alpha/(\alpha + \beta)$, and the rate of 2 → 1 polymorphisms as $\Lambda\beta/(\alpha + \beta)$; Λ is a constant proportional to the mutation rate. Using Eq. (1) these are $\Lambda(I - f)$ and Λf, respectively. Note that the overall number of new 1 → 2 mutations arising $[f\Lambda(1 - f)]$ is equal to the number of new 2 → 1 mutations $[(1 - f)\Lambda f]$. Since the two types of mutation are neutral, they have the same frequency distribution and so they are equally likely to be detected in a sample of sequences (this convenience is the sole reason for assuming the mutations are neutral).

Let us begin by considering the case where we use one outgroup sequence to infer the direction of mutation. We assume that polymorphism is relatively rare and recent so we do not have to consider polymorphism in the outgroup sequence (i.e., the outgroup sequence is assumed to be fixed at each site) and the divergence between alleles within a population is trivial compared to the divergence to the outgroup. The number of 1 → 2 and 2 → 1 mutations which we infer by parsimony are

$$
\begin{aligned}
M_{12}(1) = \Lambda\{&f(X_1[t]^2(1 - f) + X_1[t](1 - X_1[t])f)\\
&+ (1 - f)(X_2[t](1 - X_2[t])f + (1 - X_2[t])^2(1 - f))\}
\end{aligned}
\tag{10}
$$

$$
\begin{aligned}
M_{21}(1) = \Lambda\{&(1 - f)(X_2[t]^2 f + X_2[t](1 - X_2[t])(1 - f))\\
&+ f(X_1[t](1 - X_1[t])(1 - f) + (1 - X_1[t])^2 f)\}
\end{aligned}
$$

The biases are very similar to those seen with the substitution pattern (Fig. 3a). When the base composition is biased, the outgroup must be very closely related to the ingroups to correctly infer the pattern of mutation. For example, parsimony will infer that 77% of the mutations are from state 1 to state 2 when the sequences have diverged 10% (along each lineage) and the frequency of state 1 is 0.9, compared to the real figure of 50%.

Using another outgroup sequence exacerbates the problem. If we have two outgroup sequences and the three lineages split simultaneously, then

$$
\begin{aligned}
M_{12}(2) = \Lambda\{&f(X_1[t]^3(1 - f) + X_1[t]^2(1 - X_1[t])f)\\
&+ (1 - f)((1 - X_2[t])^2 X_2[t] f + (1 - X_2[t])^3(1 - f))\}
\end{aligned}
\tag{11}
$$

$$
\begin{aligned}
M_{21}(2) = \Lambda\{&(1 - f)(X_2[t]^3 f + X_2[t]^2(1 - X_2[t])(1 - f))\\
&+ f((1 - X_1[t])^2 X_1[t](1 - f) + (1 - X_1[t])^3 f)\}
\end{aligned}
$$

The proportion of mutations which are inferred to be 1 → 2 is shown in Fig. 3b; if the frequency of state 1 is 0.9 and the level of divergence 0.1, parsimony would infer that 84% of the mutations were 1 → 2, over five times as many mutations as 2 → 1.

Let us now consider the case where the lineages have not split simultaneously. Let the times of divergence be t and αt for the first and second outgroups. The equations are straightforward, but long since we now have to consider 32 trees; they are not given here, but the results are illustrated in Fig. 3c for the case where the second outgroup diverged $t/2$ generations ago. In contrast to the case of inferring the substitution pattern, the probability of correctly inferring the direction of mutation is increased if one of the lineages is quite closely related to lineage in which the polymorphism is segregating (Fig. 3c). For example, when the divergence is 0.1 (along the longest branch) and the frequency of state 1 is 0.9, only 73% of mutations are inferred to be 1 → 2 changes, compared to 84% when the sequences are all equally diverged.

Discussion

Parsimony is a simple and appealing principle by which to infer evolutionary process; however, it is known to be problematic if there are homoplasies, (i.e., parallel, back, or multiple changes) in the data. This paper examines two simple problems, inferring the pattern of substitution and inferring the pattern of mutation, from molecular data in sequences of biased base composition. Parsimony performs surprisingly badly; even with quite modest levels of sequence divergence, parsimony incorrectly infers an excess of common to rare changes. The reasons for this bias are quite simple. Consider the problem of inferring the pattern of substitution using three equally diverged sequences with two nucleotides, C and T. When sequences are stationary the number of substitutions from C to T must equal the number of substitutions from T to C. If C is more common, then the per-site probability of substitution from C to T must be lower than from T to C. Hence sites which are ancestrally T are much more likely to change to C in two lineages and hence be inferred to be C ancestrally.

The paper considers a simple two-state model, whereas there are four nucleotides in DNA. In several circumstances DNA is effectively a two-state system, for example, when the rate of transition is much higher than the rate of transversion. Furthermore, it is evident that these biases occur in real datasets; both Collins et al. (1994) and Perna and Kocher (1995) discovered this problem with parsimony when they noted the excess of common to rare changes in several DNA datasets.

The net effect of having more than two states may not be very great in many cases. Consider the problem of inferring the substitution pattern when there are three equidistant lineages. If there are substitutions at the same site in two lineages in the two-state system, we misinfer the direction and number of substitutions because both changes have to be the same. In the four-state system, this is not the case; the substitutions can be different. However, the site becomes uninformative if the changes are different, and this process is itself biased. Overall the rate at which sites with common nucleotides become uninformative will be lower than the rate for sites occupied by rare nucleotides; the inference of substitution and mutation patterns will therefore be biased.

Parsimony has been used extensively to reconstruct phylogeny from both morphological and genetic data. The current results suggest that parsimony may have problems resolving some phylogenies when the base composition is biased, since it will fail to detect some of the changes that have occured. It may be possible to overcome this problem, in part, by weighting changes according to the base composition. In a two-state system this is straightforward; if the frequency of state 1 is f, 1 → 2 changes should be weighted $(1 - f)$, 2 → 1 changes by f. For example, if the frequency of C at twofold degenerate sites is 0.9, then the probability of a T → C change is nine times higher than a C → T change; the C → T change should therefore be weighted nine fold (note that base composition can introduce much greater discrepancies between pathways than transition/transversion bias). This simple weighting scheme can be extended to the four-state case by the reducing the system to two states. For example, if the frequency of C equals the frequency of G, and A equals T, then a sensible weighting scheme would be to weight GC → AT changes by the frequency of AT and AT → GC changes by the frequency of GC. This weighting can be overlaid on other weighting schemes for transition/transversion bias.

The solution to the problems of parsimony presented here is to use more sophisticated evolutionary models which take into account biased base composition or to use other criteria to establish the direction of change (Eyre-Walker 1997). Such models have been developed for inferring the substitution pattern (Yang 1994; Yang et al. 1995; Yang and Kumar 1996), but methods have not yet been developed for inferring the direction of mutation.

Acknowledgments. I am very grateful to two anonymous referees for their comments and to the Royal Society for their support.

References

Akashi H (1995) Inferring weak selection from patterns of polymorphism and divergence at ‘‘silent’’ sites in Drosophila DNA. Genetics 139:1067–1076

Ballard JWO, Kreitman M (1994) Unravelling selection in the mitochondrial genome of Drosophila. Genetics 138:757–772

Collins TM, Wimberger PH, Naylor GJP (1994) Compositional bias, character state bias and character state reconstruction using parsimony. Syst Biol 43:482–496

Eyre-Walker A (1994) DNA mismatch repair and synonymous codon evolution in mammals. Mol Biol Evol 11:88–98

Eyre-Walker A (1997) Differentiating selection and mutation bias. Genetics 147:1983–1987

Gojobori T, Li W-H, Graur D (1982) Patterns of nucleotide substitution in pseudogenes and functional genes. J Mol Evol 18:360–369

Imanishi T, Gojobori T (1992) Patterns of nucleotide substitutions inferred from the phylogenies of the class I major histocompatability complex genes. J Mol Evol 35:196–204

Perna NT, Kocher TD (1995) Unequal base frequencies and the estimation of substitution rates. Mol Biol Evol 12:359–361

Tamura K, Nei M (1993) Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. Mol Biol Evol 10:512–526

Yang Z (1994) Estimating the pattern of nucleotide substitution. J Mol Evol 39:105–111

Yang Z, Kumar S (1996) Approximate methods for estimating the pattern of nucleotide substitution and the variation of substitution rate among sites. Mol Biol Evol 13:650–659

Yang Z, Kumar S, Nei M (1995) A new method of inference of ancestral nucleotide and amino acid sequences. Genetics 141:1641–1650