Points of View

Syst. Biol. 54(6):948–951, 2005

ISSN: 1063-5157 print / 1076-836X online
DOI: 10.1080/10635150500234682

A Tale of Two Processes

PETER LOCKHART[1] AND MIKE STEEL[2]

[1]*Institute for Molecular BioSciences, Allan Wilson Centre for Molecular Ecology and Evolution, Massey University, Palmerston North, New Zealand; E-mail: P.J.Lockhart@massey.ac.nz*

[2]*Biomathematics Research Centre, Allan Wilson Centre for Molecular Ecology and Evolution, University of Canterbury, New Zealand; E-mail: M.Steel@math.canterbury.ac.nz*

The "long-branch attraction" (LBA) phenomenon in phylogeny reconstruction is well cited but its causes have been poorly characterized. In this article, we point out that different biological processes can lead to similar forms of long-branch attraction. That is, although sequences generated by different processes look similar "through the eyes of parsimony," the ensemble of sequence site patterns (not just the parsimony sites) can distinguish between these processes.

In 1978, Felsenstein described an evolutionary scenario under which unequal amounts of change in nonadjacent lineages would mislead tree-building methods based on parsimony (or on uncorrected distances). Other authors have since shown that when substitution models are misspecified, maximum likelihood and distance-based methods can be similarly misled (e.g., Hillis et al., 1994; Lockhart et al., 1996; Bruno and Halpern, 1999; Swofford et al., 2001; Sullivan and Swofford, 2001; Ho and Jermiin, 2004). LBA problems may also arise because of sparse and/or unbalanced taxon sampling (Hendy and Penny, 1989; Holland et al. 2003; Lockhart and Penny, 2005) and/or because of lineage-specific differences in rates or processes of evolution (e.g., Hasagawa and Hashimoto, 1993; Steel et al., 1993, 2000).

Felsenstein (1978) assumed that whilst evolutionary rates varied across a tree, individual sites in sequences could be ascribed a rate of change that was the same at other sites in the same sequence. That is, if a lineage was fast (or slow) evolving, then the evolution of sites in a sequence belonging to that lineage was also fast (or slow). A site position that has evolved under this scenario can be seen as a special case of "heterotachy." This is a property of individual sequence positions, which literally means different speeds. It is the concept of sequence evolution at a given site undergoing substitution at different rates in different parts of the tree (Lopez et al., 2002). Interestingly, Simon et al. (1996) have also described this phenomenon and referred to it as "mosaic evolution." It is important to note that variation in the substitution rate of a site throughout the tree is distinct from rate variation *across* sites (as modeled, for example, by a gamma distribution). In the latter case there is a site-specific substitution rate that varies randomly across the sites, but at any site it applies equally to all the branch lengths of the tree (the branch lengths at the site are all multiplied by the site-specific rate). Consequently, the *ratio* of substitution rates on two different branches is constant across sites in such models, even when one allows both rate variation in the tree (as in Felsenstein's scenario) as well as an independent process of rate variation (e.g., gamma distribution) across sites. In contrast, with more general forms of heterotachy, the ratio of substitution rates on different branches of the tree may vary across sites.

The special case of heterotachy assumed by Felsenstein (1978) is different from another very special type of heterotachy recently explored in simulations by Kolaczkowski and Thornton (2004). These authors envisaged a four-taxon tree, for which the external lineages evolved in such a way that sites in the sequences accumulated substitutions at one of two rates, either slow or fast. As Spencer et al. (2005) point out, the frequencies of patterns expected under the simulation model studied by Kolaczkowski and Thornton (2004) are a small snapshot of the full range of possibilities when all possible combinations of short and long branches are considered, and most of these do not cause LBA (further concerns regarding the findings of Kolaczkowski and Thornton [2004] have been discussed by Steel, 2005).

Thus, the patterns that Kolaczkowski and Thornton (2004) studied are different from those expected under the standard stationary covarion (or "covarion drift") models, which have been the subject of much recent study (e.g., Tuffley and Steel, 1998; Penny et al., 2001; Gaucher et al., 2001; Huelsenbeck, 2002; Galtier, 2001; Misof et al., 2002; Inagaki et al., 2004; Ané et al., 2005; Guindon et al., 2004). These standard covarion models have reversible stationary substitution rates among character states that are switched "on" (variable), and a reversible stationary process between the state of "off" (invariable) and "on." The latter condition will maintain

the same proportion of variable sites in different lineages, though some models also allow variation in the substitution rate of sites that are "on." Under these standard covarion models, LBA might occur if faster rates of change at variable sites occur in some lineages. In this situation, tree building will be problematic in a manner similar to that previously described (Felsentein, 1978; Lockhart et al., 1996; Bruno and Halpern, 1999; Swofford et al., 2001). Evidence for covarion-like evolution in sequences has been reported by a number of authors (e.g., Fitch and Markowitch, 1970; Miyamoto and Fitch, 1995; Simon et al., 1996; Lockhart et al., 2000; Gaucher et al., 2001; Penny et al., 2001; Huelsenbeck, 2002; Misof et al., 2002; Inagaki et al., 2004; Ané et al., 2005; Brown, 2005).

The property of heterotachy might also arise in data if there is a change in the proportion of sites free to vary in different lineages. Such a property of data is presumably the basis of "covarion shifts" (Inagaki et al., 2004) and also an explanation for why, in the sequences of some taxa, there are relatively more positions with "shared-ancestral" character states than in others (Stiller et al., 2001). Concern that the proportions of variable sites can differ in biological orthologues was an initial motivation for the word "heterotachy" (Philippe and Germot 2000; Lopez et al., 2002). The phenomenon is a concern for phylogeny estimation, because if proportions of variable sites do change in different evolutionary lineages, then misleading parsimony patterns will arise similar to those that arise under the scenario described by Felsenstein (1978). Although we pointed this out previously (Lockhart et al.1998), we have only recently recognized that considering the ensemble of pattern frequencies will distinguish between the evolutionary model studied by Felsenstein (1978) and an evolutionary model in which proportions of variable sites differ between lineages. We contrast these two models as follows.

Model 1

Under this model, sequence sites evolve under a standard Markov process, with all sites evolving according to the same process. Two situations that can lead to long branch attraction are shown in Figure 1 where two nonadjacent branches undergo substitution at a markedly higher rate (1a) or where there is a long edge leading to an outgroup, and one of the two sister taxa has an elevated substitution rate (1b). However, these are just particular cases, and in Model 1 no constraints are imposed a priori on branch lengths.

Model 2

Under our second model we assume that there are two classes of sites. Proportion a evolve on the tree as in Model 1, but with branch lengths that satisfy a molecular clock. The remaining proportion $(1 - a)$ of the sites evolve on the same tree and with the same branch lengths, except that on certain branches the sites are invariable (which can be modeled by setting those particular branch lengths equal to zero). For a four-taxon tree 12 | 34, we assume that the branches on which the sites are invariable include at least one branch that connects taxa 1 and 2, and at least one branch that connects taxa 3 and 4. We also assume that a and $1 - a$ are both strictly positive (since if $a = 0$ or 1, then Model 2 is just a special case of Model 1).

A special case of Model 2 is the analogue of (1a), which is shown as the mixed process of (2a) in Figure 1, where dotted edges show branches on which the second class of sites are invariable—this also leads to long-branch attraction and causes parsimony and uncorrected distance methods to infer the same incorrect tree as (1a). Indeed, if one just considers parsimony informative patterns or pairwise distances, the models appear to be essentially similar, yet we will see shortly that consideration of the

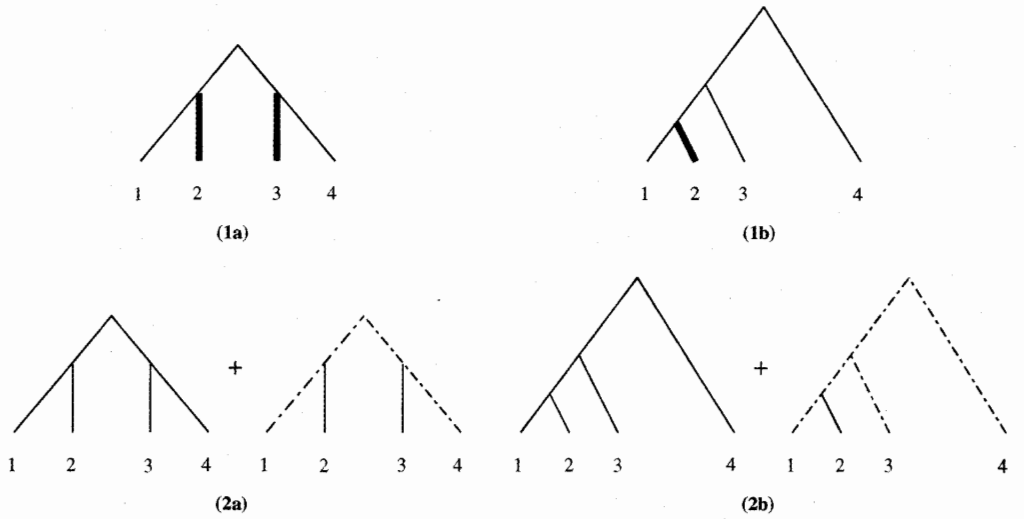

FIGURE 1. In a special case of Model 1 (the classic "Felsenstein zone") the rate of evolution is faster in two lineages (1a; bold lines) or there is a fast-evolving ingroup lineage (1b; bold line) and an outgroup lineage whose representative taxon is subtended by a long branch (1b; branch leading to 4). In Model 2, sites that are variable undergo substitution at a constant (clocklike) rate; however, some sites are "on" throughout the tree (solid edges), whereas others are invariable (dotted edges) in parts of the tree. The analogues of (1a) and (1b) for this process are shown in (2a) and (2b); however, only (2a) leads to long-branch attraction (LBA) that will mislead parsimony.

full site pattern distribution allows them to be distinguished. The Model 2 analog of (1b) is (2b) shown in Figure 1—however, in contrast to what happens in the Model 1 setting, (2b) does not lead to long-branch attraction. That is, maximum parsimony (and uncorrected distances) will converge on the correct tree for (2b) (for parsimony this follows immediately from the linearity of the parsimony-score function and the observation that the site patterns produced by the tree on the right of (2b) are not parsimony patterns; a similar simple argument applies for uncorrected distances). As with Model 1, (2a) and (2b) are just two particular instances of Model 2.

The following result shows that we can distinguish between these two models (1 or 2), provided we have enough data.

PROPOSITION

Suppose sequence sites evolve from either Model 1 or Model 2 and where variable sites evolve according to a simple (group-based) substitution models (e.g., Jukes-Cantor, Kimura 3ST). Then the model type (1 or 2) can be identified from sufficiently long sequences.

Proof

Under either model, consider the event E_{12} of a change of state between taxon 1 and 2 and the event E_{34} of change of state between taxa 3 and 4. Under Model 1 (and an underlying group-based process), it is well known (see e.g., Semple and Steel, 2003) that E_{12}, E_{34} are statistically independent, so $P[E_{12}\&E_{34}] = P[E_{12}] \times P[E_{34}]$.

In contrast, for Model 2, we claim that $P[E_{12} \& E_{34}] > P[E_{12}] \times P[E_{34}]$. To see this, note that in Model 2 the sites are in two classes—the class A_1 in which the site is variable in all branches, and the class A_2 for which the site is variable in only some of the branches. For $i = 1, 2$, we have (for group-based models)

$$P[E_{12} \& E_{34}|A_i] = P[E_{12} \mid A_i] \times P[E_{34}|A_i]$$

so that, by the law of total probability,

$$P[E_{12}\&E_{34}] = \sum_{i=1}^{2} P[E_{12}\&E_{34}|A_i]P[A_i]$$

$$= \sum_{i=1}^{2} P[E_{12}|A_i]P[E_{34}|A_i]P[A_i]$$

and we also have

$$P[E_{12}] = \sum_{i=1}^{2} P[E_{12}|A_i]P[A_i];$$

$$P[E_{34}] = \sum_{i=1}^{2} P[E_{34}|A_i]P[A_i].$$

Thus, by elementary algebra,

$$P[E_{12}\&E_{34}] - P[E_{12}]P[E_{34}] = P[A_1]P[A_2] \cdot (P[E_{12}|A_1) - P[E_{12}|A_2]) \cdot (P[E_{34}|A_1] - P[E_{34}|A_2])$$

and, by the assumptions on Model 2,

$$P[E_{12}|A_1] > P[E_{12}|A_2],\ P[E_{34}|A_1] > P[E_{34}|A_2],$$

which upon substitution into the previous equation implies that

$$P[E_{12}\&E_{34}] - P[E_{12}] \times P[E_{34}] > 0,$$

as claimed.

Because we can estimate the probabilities $P[E_{12}]$, $P[E_{34}]$, and $P[E_{12} \& E_{34}]$ arbitrarily accurately (with sufficiently long sequences) by the proportions of sites displaying these respective three patterns [p_{12}, p_{34}, and $p_{12|34}$], we can distinguish between the two models with sufficiently long sequences. This completes the proof.

Note that the ratio $p_{12}p_{34}/p_{12|34}$ is also the (Peterson) capture-recapture estimator for the proportion of variable sites under a mixed model in which a proportion of sites are variable (and evolving at a constant rate), whereas the remainder of sites are invariable (Steel et al., 2000). Curiously, although neither Model 1 nor 2 is of this type, the estimator still turns out to be useful since it is capable of distinguishing between them.

CONCLUSIONS

We have shown that it is possible in principle to distinguish between two possible causes of LBA. The simple and special case that we study is sufficient to demonstrate this principle. However, we note that our finding can also be generalized—for example, with less elegant analysis and formulae to allow an unknown proportion of invariable sites in both models. Also, it is possible to remove the restriction in Model 2 of a molecular clock; we imposed it here to keep Model 2 as simple as possible, and so that it involves no more parameters than Model 1.

Both models (1 and 2) can induce topological biases that will mislead tree building. The form identified by Felsenstein (1978) can be modeled with substitution models that are homogeneous across sites and are currently implemented in maximum likelihood and Bayesian inference software packages. However, a process whereby the proportion of variable sites change across the underlying phylogeny is not currently modeled (Inagaki et al., 2004). Such a model would involve either a mixture of different processes or a nonstationary process (such nonstationary processes have previously been applied to model the evolution of sequences exhibiting base composition variation).

In practice, distinguishing empirical examples where the proportion of variable sites changes from cases where the process of sequence evolution is described by the standard covarion model is complicated by the fact that the latter model can appear to indicate an increase in the proportion of variable sites in groups with larger evolutionary divergence. Nevertheless, we have recently attempted to make an objective test of whether or not proportions of variable sites change among orthologues, and we have reported observations suggesting

lineage-specific differences among eubacterial and plastid sequences (Lockhart et al., submitted). In support of this inference we have also noted lineage-specific differences in the occurrence of indels and nonconservative substitutions, consistent with the relaxation of functional/structural constraints in lineages that have also increased their proportion of variable sites. These observations provide an indication of the importance of the relationship between sites that are free to vary in sequences and evolving evolutionary constraints. However, more detailed empirical studies are needed to investigate the degree to which heterotachy is nonhomogenous in nature. Further, the extent to which (a) homogenous maximum likelihood models can accommodate heterotachy in real data and (b) the extent to which nonreversible biological processes can be effectively modeled also requires further study. In respect to (a), the empirical findings that we report in Lockhart et al. (2005) are encouraging and are in contrast to the findings from simulations, reported by Kolaczkowski and Thornton (2004).

REFERENCES

Ané, C., J. G. Burleigh, M. M. McMahon, and M. J. Sanderson. 2005. Covarion structure in plastid genome evolution: A new statistical test. Mol. Biol. Evol. 22:914–924.

Brown, R. P. 2005. Large subunit mitochondrial rRNA secondary structures and site-specific rate variation in two lizard lineages. J. Mol. Evol. 60:45–56.

Bruno, W. J., and A. L. Halpern. 1999. Topological bias and inconsistency of maximum likelihood using wrong models. Mol. Biol. Evol. 16:564–566.

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

Fitch, W. M., and E. Markowitz. 1970. An improved method for determining codon variability in a gene and its application to the rate of fixation of mutations in evolution. Biochem. Genet. 4:579–593.

Galtier, N. 2001. Maximum-likelihood phylogenetic analysis under a covarion-like model. Mol Biol. Evol. 18:866–873.

Gaucher, E. A., M. M. Miyamoto, and S. A. Benner. 2001. Function-structure analysis of proteins using covarion-based evolutionary approaches: Elongation factors. Proc. Natl. Acad. Sci. USA 98:548–552.

Guindon, S., A. G. Rodrigo, K. A. Dyer, and J. P. Huelsenbeck. 2004. Modelling the site-specific variation of selection patterns along lineages. Proc Natl. Acad. Sci. USA 101:12957–12962.

Hasegawa, M., and T. Hashimoto. 1993. Ribosomal RNA trees misleading? Nature 361:23.

Hendy, M. D., and D. Penny. 1989. A framework for the quantitative study of evolutionary trees. Syst. Zool. 38:297–309.

Hillis, D. M., J. P. Huelsenbeck, and D. L. Swofford. 1994. Hobglobin of phylogenetics? Nature 369:363–364.

Ho, S.Y., and L. Jermiin. 2004. Tracing the decay of the historical signal in biological sequence data. Syst. Biol. 53:623–637.

Holland, B. R., D. Penny, and M. D. Hendy. 2003. Outgroup misplacement and phylogenetic inaccuracy under a molecular clock—a simulation study. Syst. Biol. 52:229–238.

Huelsenbeck, J. P. 2002. Testing a covariotide model of DNA substitution. Mol. Biol. Evol. 19:698–707.

Inagaki, Y., E. Susko, N. M. Fast, and A. J. Roger. 2004. Covarion shifts cause a long-branch attraction artefact that unites Microsporidia and Archaebacteria in EF-1α phylogenies. Mol. Biol. Evol. 21:1340–1349.

Kolaczkowski, B., and J. W. Thornton. 2004. Performance of maximum parsimony and likelihood phylogenetics when evolution is heterogeneous. Nature 431:980–984.

Lockhart, P. J., D. Huson, U.-G. Maier, M. J. Faunholz, Y. Van de Peer, A. C. Barbrook, C. J. Howe, and M. A. Steel. 2000. How molecules evolve in eubacteria. Mol. Biol. Evol. 17:835–838.

Lockhart, P. J., A. W. D. Larkum, M. A. Steel, P. J. Waddell, and D. Penny. 1996. Evolution of chlorophyll and bacteriochlorophyll: The problem of invariant sites in sequence analysis. Proc. Natl. Acad. Sci. USA 93:1930–1934.

Lockhart, P. J., P. Novis, B. G. Milligan, J. Riden, A. Rambaut, and A. W. D. Larkum. 2005. Heterotachy and tree building: A case study with plastids and eubacteria Mol. Biol. Evol. in press.

Lockhart, P. J., and D. Penny. 2005. The place of *Amborella* in the radiation of angiosperms. Trends in Plant Sci. 10:201–202.

Lockhart, P. J., M. A. Steel, A. C. Barbrook, D. H. Huson, and C. J. Howe. 1998. A covariotide model describes the evolution of oxygenic photosynthesis. Mol. Biol. Evol. 15:1183–1188.

Lopez, P., D. Casane, and H. Philippe. 2002. Heterotachy, an important process of protein evolution. Mol. Biol. Evol. 19:1–7.

Misof, B., C. L. Anderson, T. R. Buckley, D. Erpenbeck, A. Rickert, and K. Misof. 2002. An empirical analysis of mt 16S rRNA covarion-like evolution in insects: Site-specific rate variation is clustered and frequently detected. J. Mol. Evol. 55:460–469.

Miyamoto, M. M., and W. M. Fitch. 1995. Testing the covarion hypothesis of molecular evolution. Mol. Biol. Evol. 12:503–513.

Penny, D., B. J. McComish, M. A. Charleston, and M. D. Hendy. 2001. Mathematical elegance with biochemical realism: The covarion model of molecular evolution. J. Mol. Evol. 53:711–723.

Philippe, H., and A. Germot. 2000. Phylogeny of eukaryotes based on ribosomal RNA: long-branch attraction and models of sequence evolution. Mol. Biol. Evol. 17:830–834.

Philippe, H., and P. Lopez. 2001. On the conservation of protein sequences in evolution. Trends Biochem. Sci.26:414–416.

Semple, C., and M. Steel. 2003. Phylogenetics. Oxford University Press, Oxford, UK.

Simon, C., L. Nigro, J. Sullivan, K. Holsinger, A. Martin, A. Grapputo, A. Franke, and C. McIntosh. 1996. Large differences in substitutional pattern and evolutionary rate of 12S ribosomal RNA genes. Mol. Biol. Evol. 13:923–932.

Spencer, M., E. Susko, and A. J. Roger. 2005. Likelihood, parsimony, and heterogeneous evolution. Mol. Biol. Evol. 22:1161–1164.

Steel, M. A. 2005. Should phylogenetic models be trying to 'fit an elephant?' Trends Geneti. 21:307–309.

Steel, M. A., D. Huson, and P. J. Lockhart. 2000. Invariable site models and their use in phylogeny reconstruction Syst. Biol. 49:225–232.

Steel, M. A., P. J. Lockhart, and D. Penny. 1993. Confidence in evolutionary trees from biological sequence data. Nature 364:440–442.

Stiller, J. W., J. Riley, and B. D. Hall. 2001. Are Red Algae Plants? A critical evaluation of three key molecular data sets. J. Mol. Evol. 2001. 52:527–539.

Sullivan, J. A., and D. L. Swofford. 2001. Should we use model-based methods for phylogenetic inference when we know that assumptions about among-site rate variation and nucleotide substitution pattern are violated? Syst. Biol. 50:723–729.

Swofford, D. L., P. J. Waddell, J. P. Huelsenbeck, P. G. Foster, P. O. Lewis, and J. S. Rogers. 2001. Bias in phylogenetic estimation and its relevance to the choice between parsimony and likelihood methods Syst. Biol. 50:525–539.

Tuffley, C., and M. Steel. 1998. Modeling the covarion hypothesis of nucleotide substitution. Math. Biosci. 147:63–91.

First submitted 13 March 2005; reviews returned 21 April 2005; final acceptance 3 May 2005

Associate Editor: Jack Sullivan