Update TRENDS in Genetics Vol.21 No.6 June 2005

Full text provided by www.sciencedirect.com

Research Focus

Should phylogenetic models be trying to 'fit an elephant'?

Mike Steel

Allan Wilson Centre for Molecular Ecology and Evolution, Biomathematics Research Centre, University of Canterbury, Christchurch, New Zealand

For the past two decades, there has been an ongoing debate within the plylogenetics community over whether model-based approaches for molecular systematics (such as maximum likelihood) should be preferred over the more traditional 'maximum parsimony' approach. A recent simulation study by Kolaczkowski and Thornton has brought this debate into sharp focus. In this article, I discuss the significance of their findings and offer a prognosis on the implications for molecular phylogenetics. I believe that biochemistry and model selection have an important role in developing accurate phylogenetic approaches.

The ongoing debate concerning model-based phylogenetics

Nucleic acid sequences carry with them information about the evolutionary relationships of their host species, and phylogenetics techniques try to extract this historical 'signal' (Box 1).

During the past twenty years, the use of stochastic models has completely transformed the field of phylogenetics [1]. Maximum likelihood (ML) and, more recently, bayesian and model-selection approaches, such as Akaike information criterion (AIC) and bayesian information criterion (BIC), have become the methods of choice for most molecular studies. However, a vocal resistance to these approaches remains, with the claim that model-based approaches are inherently flawed owing to the unrealistic nature of current models or the underlying problems in determining an accurate model. Advocates of this viewpoint often argue that maximum parsimony (MP) provides the only sound basis for the phylogenetic analysis of sequence data. To support their position, various articles claim to show that ML can perform poorly compared with MP on certain simulated data [2,3].

The findings of Kolaczkowski and Thornton

The recent article by Kolaczkowski and Thornton [4] sheds some interesting light on this debate. Their investigation was motivated by recent studies suggesting that heterotachy – changes in the substitution rate of certain sequence sites at various locations in the phylogenetic tree – might be an important feature of sequence evolution [5]. To model this, they considered a simple scenario in which half of the sequence sites evolve under one set of substitution rates and half evolve under another set, on a four-taxon phylogenetic tree and according to the simplest substitution process (i.e. the Jukes–Cantor process). They allowed half of the sites to evolve with elevated substitution rates on two non-adjacent terminal branches (Figure 1a), whereas the remaining sites evolve under a 'mirror-image' process – where the fast and slow rates on the terminal branches are interchanged (Figure 1b). When the length of the internal branch decreases to zero most methods will favour the incorrect phylogenetic tree that combines the non-adjacent taxa that share their matching branch lengths (ad and bc). Kolaczkowski and Thornton found that MP tends to 'outperform' ML and a bayesian approach when those methods are used in a variety of models (using standard models that do not include the type of model that generated the data). Here 'outperform' means that MP remains more accurate at shorter interior-branch lengths than either ML or bayesian methods.

When Kolaczkowski and Thornton used a bayesian analysis based on the type of model that generated the data, they found that this bayesian Markov Chain Monte Carlo (BMCMChetero) method was more accurate than either MP or ML. This bayesian method allowed two classes of sequence sites, with the sites in each class being

Box 1. A brief overview of phylogenetics and its terminology

A phylogenetic study usually begins with the alignment of nucleic acid sequences – one for each species in a collection of species that are being studied; at each site (position) in the sequence, each species has one of four (DNA or RNA) or 20 (amino acid) states. These site patterns convey information about the history of the species – in particular the phylogenetic tree (branching order) of speciation and the divergence times of different lineages. Sequences that are identical at most sites suggest that the species they are sampled from are closely related, and therefore close to each other in the phylogenetic tree, whereas more dissimilar sequences suggest greater evolutionary separation. However, the relationship between evolutionary distance (distance in the tree) and sequence dissimilarity is not linear and other complications arise: for example, the rate of substitution can vary across the tree and across the sequence sites. Sequence-based phylogenetic methods are usually based either on some non-parametric optimality criteria or on an explicit model of nucleic acid substitution. Maximum parsimony is a non-parametric method and it seeks the tree(s) requiring the fewest substitution events to account for the observed data. Methods based on an explicit model of nucleic acid substitution include maximum likelihood (ML), bayesian methods and model selection (e.g. AIC and BIC) approaches.

Corresponding author: Steel, M. (m.steel@math.canterbury.ac.nz).

Available online 12 April 2005

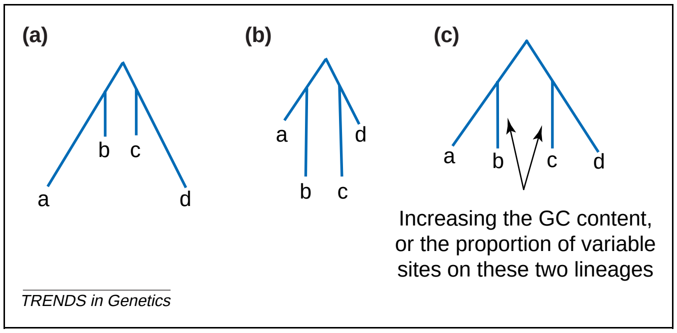
(a)
b c
a
d
(b)
a
d
b c
(c)
a
b
c
d
Increasing the GC content, or the proportion of variable sites on these two lineages
TRENDS in Genetics

Figure 1. Mechanisms that cause problems for certain tree-reconstruction methods. The two opposing sets of branch lengths **(a)** and **(b)** were used in the simulation of Kolaczkowski and Thornton [4]. **(c)** A change in the process which can mislead all methods (MP, ML) that assume molecular evolution is governed by a common process across the tree. All trees are drawn as rooted to emphasize fit and departure from a constant substitution rate (molecular clock) at the variable sites.

governed by the same branch length setting, but these two settings were allowed to differ (neither the classes nor the branch lengths were assumed *a priori* but were estimated from the data). This result shows that ML approaches to heterotachy might be promising in the future but, as the authors note, these approaches raise certain issues. For example, with real-world sequence data, it might be difficult to know *a priori* how many site classes would be needed to represent data accurately, and one might go to the extreme of allowing each site to evolve under its own particular suite of branch lengths. In this extreme case, it is known [6] that the ML tree(s) for the Jukes–Cantor process is exactly the same as MP tree(s), the implications of which were recently discussed by Sober [7]. The authors also noted other potential issues associated with complex, highly parameterized models, in particular the computational burden of finding optimal trees. In summary, Kolaczkowski and Thornton recommend 'interpreting likelihood-based inferences with the same caution that is currently applied to maximum parsimony'.

Consequences of model mis-specification

It is generally accepted that most models in molecular systematics are overly simplistic, and that model mis-specification can mislead model-based tree-reconstruction methods. The inability of ML to resolve the correct tree in Kolaczkowski and Thornton's article [4] at certain parameter settings is not surprising. ML can be statistically inconsistent (fail to converge on the true tree with increasing data) when the data are generated (by nature or by computer simulation) according to a mechanism that differs from those used in the likelihood analysis [8]. Inconsistency can also arise even when the true model is used if that model is too parameter-rich, a problem known as 'non-identifyability' [9]. Moreover, MP can outperform ML even when the underlying model is correct, provided the tree and parameters are chosen correctly (e.g. a four-taxon tree with three long branches and two short, adjacent terminal branches). This phenomenon has been noted by several authors, most vocally by Siddall [2], although some claims in that article were subsequently refuted by Swofford *et al.* [10].

Biochemical realism

Kolaczkowski and Thornton point out that the hypothetical mixture model used in their simulation is a simplified model; therefore, is it relevant for real sequence data? The type of heterotachy that they described seems biologically implausible – what possible biochemical mechanism would suggest that the rates on the four branches between the two classes of sites are anti-correlated in the way that Figure 1a,b suggests? This pattern is improbable under biochemically motivated models that describe how the substitution process might vary across a tree. Simple biochemical mechanisms have been discussed that can give the same type of misleading influence – for both ML and MP – as the artificial process described by Kolaczkowski and Thornton. In the classic 'long-branch attraction' setting [11], the two 'long' branches can be viewed as the combination of a long branch (to an outlying taxon) with a rate acceleration in one lineages or, perhaps, as two lineages where the rate of substitution has been independently elevated. However, there are processes that lead to similar long-branch attraction, even when the substitution rate at each variable site is constant across the tree. Two such processes involve a change in the substitution process on two non-adjacent edges (Figure 1c). The first process is an increase in the GC-content of the two lineages [12–15]. The second process is a change in the proportion and/or distribution of variable sites in these two lineages [16–18], presumably because of structural or functional reasons ('covarion-type' models). It seems that both of these processes occur in certain data sets, and they might be the main cause of 'long branch attraction' (Peter Lockhart personal communication).

Model fitting and the 'elephant' factor

Kolaczkowski and Thornton's findings can be viewed as supporting a likelihood approach over MP because an ML analysis performed under the correct (two-process) model outperforms MP. However, as the authors ask, how can one know in advance whether this model is the correct one to use and therefore avoid adding more categories of sites and parameters?

Two approaches might be useful to answer this question. First, the sequence data or biochemical information (e.g. functional or structural properties of the sequences or DNA-repair mechanisms) often provides important clues to which processes were involved in sequence evolution. For example, in coding sequences, models that respect the greater redundancy in the third codon position are more effective than those that treat all sites equally. The empirical frequencies of different amino acids at sites can also be exploited in models [19], as can variations in the empirical base frequencies (GC-content) between groups in the tree. In the second case, a model that assumes a constant substitution process across the tree is inappropriate, and a more accurate model would incorporate compositional variation. This can lead to new methodology (e.g. the 'logdet' transform) and help explain why some data produce misleading phylogeny using standard methods [13]. Numerous models have since been developed to explain some of the intricacies of sequence evolution [19–21], although comparatively little work has been performed on heterotachy, beyond simple covarion-type models; therefore, Kolaczkowski and Thornton's article will no doubt help encourage further investigation.

The second consideration in choosing a model is a well-developed statistical theory (model selection and model averaging) that enables models to be compared according to objective criteria, such as AIC or BIC. Briefly, these approaches penalise the addition of extra parameters, unless there is a sufficiently impressive improvement in fit between model and data. Unlike traditional ML ratio tests these approaches enable the comparison of any two models, even if they are not nested [22]. Allowing progressively more parameters always leads to an improvement of fit between the data and the model – a phenomenon that has become a folklore quote (from physics): 'with enough parameters you can fit an elephant'. However, the extensive addition of parameters comes at a price – the predictive power of the theory (the information that the data can reveal about the underlying tree) tends to be drowned out in a sea of parameter estimation. The aim of model selection is not to find the 'true model' but to find a model with sufficient parameters to capture the key features of the data, including the historical signal.

Concluding remarks

In summary, 'better, more realistic models' should not mean 'more parameter-rich models' – these might 'capture' more of reality, but only when the numerous parameters that are required are close to their correct values. However, the power of a given amount of data to estimate several parameters accurately is generally low. Modest parameter models that capture the main features of the sequence data are more useful – learning how DNA evolves is crucial to this task and a challenge for the future. As Simon Tavaré used to urge us many years ago: 'talk to the biochemists!'

Acknowledgements

I thank Peter Lockhart for helpful comments, and the New Zealand Marsden Fund for support. I also thank the referees and Joe Thornton for many constructive comments.

References

1 Felsenstein, J. (2004) *Inferring Phylogenies*, Sinauer Press
2 Siddall, M.E. (1998) Success of parsimony in the four-taxon case: long-branch repulsion by likelihood in the Farris zone. *Cladistics* 14, 209–220
3 Pol, D. and Siddall, M.E. (2001) Biases in maximum likelihood and parsimony: a simulation approach to a ten-taxon case. *Cladistics* 17, 266–281
4 Kolaczkowski, B. and Thornton, J.W. (2004) Performance of maximum parsimony and likelihood phylogenetics when evolution is heterogeneous. *Nature* 431, 980–984
5 Lopez, P. *et al*. (2002) Heterotachy, an important process of protein evolution. *Mol. Biol. Evol.* 19, 1–7
6 Tuffley, C. and Steel, M.A. (1997) Links between maximum likelihood and maximum parsimony under a simple model of site substitution. *Bull. Math. Biol.* 59, 581–607
7 Sober, E. (2004) The contest between parsimony and likelihood. *Syst. Biol.* 53, 644–653
8 Chang, J.T. (1996) Inconsistency of evolutionary tree topology reconstruction methods when substitution rates vary across characters. *Math. Biosci.* 134, 189–215
9 Chang, J.T. (1996) Full reconstruction of Markov models on evolutionary trees: identifyability and consistency. *Math. Biosci.* 137, 51–73
10 Swofford, D.L. *et al*. (2001) Bias in phylogenetic estimation and its relevance to the choice between parsimony and likelihood methods. *Syst. Biol.* 50, 525–539
11 Felsenstein, J. (1978) Cases in which parsimony or compatibility methods will be positively misleading. *Syst. Zool.* 27, 401–410
12 Lockhart, P.J. *et al*. (1992) Substitutional bias confounds inference of cyanelle origins from sequence data. *J. Mol. Evol.* 34, 153–162
13 Lockhart, P.J. *et al*. (1994) Recovering the correct evolutionary tree under a more realistic model of sequence evolution. *Mol. Biol. Evol.* 11, 605–612
14 Steel, M.A. *et al*. (1993) Confidence in evolutionary trees from biological sequence data. *Nature* 364, 440–442
15 Jermiin, L.S. *et al*. (2004) The biasing effect of compositional heterogeneity on phylogenetic estimates may be underestimated. *Syst. Biol.* 53, 638–643
16 Lockhart, P.J. *et al*. (1998) A covariotide model describes the evolution of oxygenic photosynthesis. *Mol. Biol. Evol.* 15, 1183–1188
17 Steel, M.A. *et al*. (2000) Invariable site models and their use in phylogeny reconstruction. *Syst. Biol.* 49, 225–232
18 Inagaki, Y. *et al*. (2004) Covarion shifts cause a long-branch attraction artefact that unites Microsporidia and Archaebacteria in EF-1α phylogenies. *Mol. Biol. Evol.* 21, 1340–1349
19 Halpern, A.L. and Bruno, W.J. (1998) Evolutionary distances for protein-coding sequences: modelling site-specific residue frequencies. *Mol. Biol. Evol.* 15, 910–917
20 Jensen, J.L. and Pedersen, A-M.K. (2000) Probabilistic models of DNA sequence evolution with context dependent rates of substitution. *Adv. Appl. Probab.* 32, 499–517
21 Galtier, N. and Gouy, M. (1998) Inferring pattern and process: maximum-likelihood implementation of a nonhomogeneous model of DNA sequence evolution for phylogenetic analysis. *Mol. Biol. Evol.* 15, 871–879
22 Posada, D. and Buckley, T. (2004) Model selection and model averaging in phylogenetics: advantages of akaike information criterion and bayesian approaches over likelihood ratio tests. *Syst. Biol.* 53, 793–808

doi:10.1016/j.tig.2005.04.001