Among-site rate variation and its impact on phylogenetic analyses

Ziheng Yang

Although several decades of study have revealed the ubiquity of variation of evolutionary rates among sites, reliable methods for studying rate variation were not developed until very recently. Early methods fit theoretical distributions to the numbers of changes at sites inferred by parsimony and substantially underestimate the rate variation. Recent analyses show that failure to account for rate variation can have drastic effects, leading to biased dating of speciation events, biased estimation of the transition:transversion rate ratio, and incorrect reconstruction of phylogenies.

Ziheng Yang is at the Dept of Integrative Biology, University of California, Berkeley, CA 94720, USA (ziheng@mws4.biol.berkeley.edu), and the College of Animal Science and Technology, Beijing Agricultural University, Beijing 100094, China.

Beginning in the 1960s[1], evolutionary studies have revealed that substitution rate variation exists among sites in almost all genes or proteins, with the possible exception of some pseudogenes or 'junk' DNA. Although mutation rates may vary among sites[2], the major reason for the variation of evolutionary rates appears to be different selective constraints at different sites owing to the functional and/or structural requirements of the gene or protein. For example, estimated substitution rates at the first (r_1), second (r_2) and third (r_3) codon positions are almost always in the order $r_2 < r_1 < r_3$, and proteins performing fundamentally important roles tend to evolve more slowly than other proteins (e.g. see Ref. 3).

If all sites in a sequence change at the same rate, the number of substitutions per site for a group of sequences should follow a Poisson distribution. Fitch and Margoliash[1] counted the minimum number of nucleotide changes at each site in cytochrome *c*, and found that the Poisson distribution did not fit the data unless a certain number of 'invariant' and 'hypermutable' sites were excluded. Their analysis effectively used three classes of rates for sites.

Continuous distributions are also used to model rate variation among sites, and by far the most-commonly used continuous distribution is the gamma distribution[4–7]. When rates at sites are gamma distributed, the numbers of substitutions at sites should follow a negative binomial distribution. Golding[5] provided an excellent summary of early studies that employ this idea. A number of formulae have also been suggested for estimating the sequence divergence under the gamma model, enabling the rate variation to be taken into account in distance-based phylogenetic analyses[5,8–12]. These methods, however, require the parameter of the gamma distribution to be specified or independently estimated.

Early studies of rate variation attempted to fit theoretical distributions to the numbers of changes at sites inferred by the parsimony method, and suffer from the systematic errors introduced in the inference (see below). When more reliable methods were developed, it became clear that these parsimony-based methods considerably underestimate the extent of among-site rate variation.

Models for variable substitution rates among sites

The standard approach to characterizing among-site rate variation is to use a statistical distribution, either discrete or continuous, to approximate rates at sites. The working hypothesis is that each site has an (unknown) rate that is determined by its position in the molecule; a fast-changing site is assumed to experience substitutions at an elevated rate in all evolutionary lineages no matter what nucleotide occupies the site.

The discrete-distribution model assumes that the rate for a site comes from one of several rate classes. With *K* classes, the model involves $2(K-1)$ free parameters (K frequency parameters with their sum equal to one and K rate parameters with the mean rate equal to one). The model is generally used with two or three rate classes only. The simplest but also the most-frequently used model of this nature is an 'invariable-sites model', which assumes that a proportion of sites have rate zero, while other sites change at the same rate[13–15]. A major problem with this model is that estimates of the proportion of invariable sites depend to a large extent on the number and relatedness of sequences (taxa) in the data, indicating the inadequacy of the model; ideally this parameter should reflect the extent of rate variation.

Based on biological considerations, one should expect a continuum of rates at sites[14]. The most-commonly used

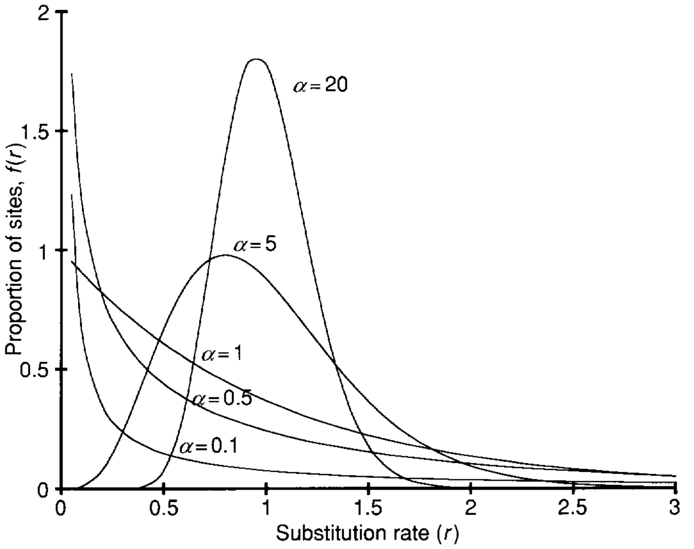


Fig. 1. The density function, $f(r)$, of the gamma distribution of substitution rates at sites (r). The gamma distribution has a shape parameter α and a scale parameter β, with mean α/β and variance α/β^2. Since the rate is a proportional factor, β is fixed to be equal to α so that the mean of the distribution is 1 and the variance is $1/\alpha$. The single parameter α is then inversely related to the extent of rate variation. The distribution with $\alpha \leq 1$ is L-shaped, meaning that most sites have very low substitution rates or are virtually 'invariable', while a few sites exist (substitutional 'hot spots') with very high rates. The distribution with $\alpha > 1$ is bell-shaped, meaning that most sites have intermediate rates while few sites have very low or very high rates. When α approaches ∞, the model reduces to the case of a constant rate for all sites. By adjusting α, the gamma model can account for different levels of rate variation in real data.

0169-5347/96/$15.00 PII: S0169-5347(96)10041-0

Box 1. Models of rate variation – advantages and disadvantages

Discrete rate-class models

Advantages

Calculations involved in the maximum likelihood analysis are relatively simple and fast.

Disadvantages

Estimates of parameters such as the proportion of invariable sites are sensitive to sampling of sequences (taxa), making the interpretation of the model difficult. Results are not comparable across data sets or analyses if different numbers of classes are used. For most data sets, two rate classes are not enough to achieve an adequate fit while three rate classes introduce too many parameters to be estimated.

Continuous rate models (gamma distribution)

Advantages

One parameter completely describes the rate variation and the model is easy to interpret. The model provides a good fit to many data sets, most often better than a two-rate-class model, which involves one more parameter than the gamma model.

Disadvantages

The likelihood calculation involves intensive computation and is feasible only for small data sets (no more than six sequences).

Discrete-gamma model

Advantages

By using well-chosen rate classes to approximate the continuous gamma, this model appears to have both the easy interpretability and good fit of the continuous gamma model and the computational efficiency of the discrete rate-class model.

continuous distribution for modelling the rate variation is the gamma distribution, although the log-normal is used occasionally. The gamma distribution involves a 'shape' parameter α (>0), which determines the extent of rate variation among sites, with a small α representing extreme rate variation and a large α representing minor rate variation (Fig. 1). The distribution is either bell-shaped or L-shaped depending on whether or not α is greater than one; this makes the distribution suitable for accommodating different levels of rate variation in various data sets. Some workers also added a proportion of invariable sites to the gamma model[14,16]. As discussed by Golding[5], this may not be worthwhile as the gamma distribution is sufficiently general to allow for very low rates at some sites.

The advantages and disadvantages of the discrete rate-class model versus the continuous gamma model are summarized in Box 1. The gamma model often fits data better than a two-rate-class model[6,17,18], but not as well as a three-rate-class model[17]. However, the gamma model involves one parameter while the two- and three-rate-class models involve two and four parameters, respectively. The gamma model appears preferable owing to its easy interpretability and good fit to many data sets (Box 1).

Yang[17,19] developed models that account for rate differences among genes and rate variation within the same gene. These models account for the heterogeneity of different genes and are useful for combined analysis of multiple sequence data. Models that allow for the correlation of substitution rates at adjacent sites were also considered[17,20]. A strong positive correlation is found in the analyzed data; this does not seem to affect parameter estimates greatly but affects the calculation of their variances[17]. Gu *et al.*[16] extended the maximum likelihood gamma model[7] to include invariable sites, and Kelly and Rice[21] investigated the problem without assuming a specific distribution. Since phylogenetic analyses appear more or less robust to the form of the distribution for the rates[17], these models are expected to produce results similar to the simple gamma model.

Estimation of the α parameter of the gamma distribution

If substitution rates are identical among sites, the numbers of substitutions at sites should follow a Poisson distribution, and if rates are gamma distributed, the numbers of changes will follow a negative binomial distribution. This principle has been used to estimate the gamma parameter, α, and to test for rate constancy among sites, with the method of maximum parsimony used to infer the (minimum) numbers of changes at sites. The α parameter is usually calculated by equating the mean and variance of the inferred numbers of changes at sites to the mean and variance of the negative binomial. This 'method of moments' estimator was used extensively for estimating α until only a few years ago[5,9]. Recently, Sullivan *et al.*[22] noted that the method of moments will overestimate α when α is small, and suggested the use of a maximum likelihood criterion to fit the negative binomial. The method was further improved by Yang and Kumar[12], who corrected for multiple hits within branches of the phylogenetic tree under the gamma model. These methods are

Table 1. Maximum likelihood estimates of the α parameter[a]

Sequences	Species	$\hat{\alpha}$	Refs
Nuclear genes			
α- and β-globin genes, positions 1 and 2	5 mammals	0.36	10,23
Albumin genes, all positions	5 vertebrates	1.05	44
Insulin genes, all positions	5 vertebrates	0.40	44
c-myc genes, all positions	5 vertebrates	0.47	44
Prolactin genes, all positions	5 vertebrates	1.37	44
16S-like rRNAs, stem region	5 species	0.29	45
16S-like rRNAs, loop region	5 species	0.58	45
$\psi\eta$-globin pseudogenes	6 primates	0.66	23
Viral genes			
Hepatitis B virus genomes	13 variants	0.26	46
Mitochondrial genes			
12S rRNAs	9 rodents	0.16	22
895-bp mtDNAs	9 primates	0.43	10
Positions 1 and 2 of 13 genes[b]	11 vertebrates	0.13–0.95	28
Position 1 of four genes	6 primates	0.18	19
Position 2 of four genes	6 primates	0.08	19
Position 3 of four genes	6 primates	1.58	19
D-loop region of mtDNAs[c]	25 humans	0.17	12
Protein sequences			
Mitochondrial cytochrome *b*	16 deuterostomes	0.44	12

[a] These estimates are all obtained from the maximum likelihood analyses of the original sequence data[7,10]. For nucleotide sequences, the assumed substitution model[14,34] accounts for both the transition:transversion rate bias and nucleotide frequency differences. For amino acid sequences, the empirical model of Jones *et al.*[47] is used.

[b] The 13 protein-coding genes in the mitochondrial genome are analyzed separately, with only the first and second codon positions used. The estimates of α are 0.49, 0.86, 0.13, 0.33, 0.23, 0.23, 0.27, 0.66, 0.39, 0.45, 0.87, 0.46 and 0.95, for *Atp6*, *Atp8*, *Cox1*, *Cox2*, *Cox3*, *Cytb*, *Ndh1*, *Ndh2*, *Ndh3*, *Ndh4*, *Ndh4L*, *Ndh5* and *Ndh6*, respectively[28].

[c] The data contain the two hypervariable segments but not the middle segment, which is almost invariable[12].

computationally fast, even for very large data sets. All three methods, however, are based on the numbers of changes at sites inferred by parsimony, which are clearly underestimates. The real problem is that the underestimation is more serious at fast-changing sites than at slowly-changing sites. This causes the three parsimony-based methods to underestimate the extent of rate variation and overestimate α, especially for distantly related sequences.

A method that avoids these problems is the maximum likelihood method applied to the original sequence data[7]. By using an explicit model of nucleotide substitution, this method simultaneously accounts for transition/transversion rate bias, unequal nucleotide frequencies, and rate variation among sites. However, the method involves intensive computation and is presently feasible for small data sets only. A more-practical approach is the 'discrete-gamma' model[10], in which several classes of rates are used to approximate the continuous gamma distribution. The classes are chosen such that each has equal probability and the mean of the rates included in a class is used to represent all rates in that class. This model combines the computational efficiency of the rate-class model with the good fit of the continuous gamma model (Box 1). With current computing power, this method is feasible for analyzing data sets of over 50 sequences.

All methods for estimating α discussed above rely on a phylogenetic tree. The effect of tree topology was also examined in several studies[10,17,23,24]. It was found that if the used tree is completely wrong, α will be seriously underestimated. Nevertheless, estimates obtained using reasonable trees (for example, trees that maintain well-supported partitions separated by long internal branches[24]) are quite similar. The relative stability of estimates of α across reasonable topologies means that reliable estimates of α are obtainable from real data even if the phylogeny of the species is uncertain.

It is also worthwhile to estimate the substitution rate for each site. Intuitive methods for examining site variability along the sequence calculate a 'conservation' or 'variability' score for each site and plot the score along the sequence using certain smoothing algorithms. These methods do not use any phylogenetic information but appear to be powerful in identifying conserved and variable regions in a DNA or protein sequence. Methods[17,25] that take into account the phylogenetic relationship of the sequences can be expected to produce more accurate estimates of substitution rates at sites, which are often found to correlate with the functional domains of the gene[12,26].

Table 1 lists estimates of α, obtained from maximum likelihood analyses of various data sets. Many estimates of α were obtained using parsimony during the past two or three decades (see Ref. 5) but these are not listed here as they are likely to be unreliable (see below). It is noticeable that most estimates are <1 so that the distribution of rates is L-shaped (Fig. 1). Larger estimates are obtained for pseudogenes or the third codon position, where substitution rates are more or less homogeneous across sites.

In Table 2, estimates of α obtained using parsimony-based methods are compared with the more-reliable likelihood estimates. The three parsimony methods give substantial overestimates of α, although the two more-recent methods[12,22] are able to reduce the bias considerably. In other analyses, parsimony (method of moments) was found to overestimate α by five or eight times relative to the maximum likelihood estimates[27,28]. Besides the estimation methods, the number of sequences (taxa) is also an important factor affecting the accuracy of the estimated α.

Table 2. Estimates of the α parameter by different methods

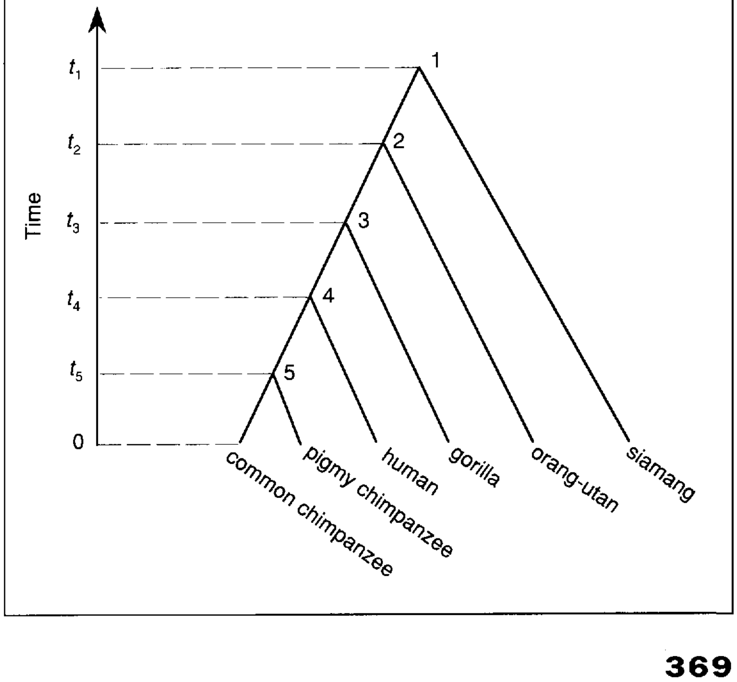
	Maximum parsimony			Maximum likelihood
Data	*a*	*b*	*c*	*d*
D-loop mtDNAs of 25 humans	0.26	0.23	0.18	0.17
16S-like rRNAs of 17 eukaryotes	0.57	0.45	0.37	0.31
mt cytochrome *b* of 16 deuterostomes	0.93	0.61	0.55	0.44

The maximum parsimony methods infer the (minimum) numbers of changes at sites, and fit theoretical expectations to these numbers by the method of moments (*a*), the likelihood method of Sullivan *et al.*[22] (*b*), or the likelihood method of Yang and Kumar[12] (*c*). In a maximum likelihood analysis (*d*), the discrete-gamma model[10] is fitted to the original sequence data (rather than to the inferred numbers of changes at sites) in combination with a model of nucleotide or amino acid substitution. The estimates are expected to be in the order *a* > *b* > *c* > *d*, with *d* being the most reliable. *From Ref. 12.*

Box 2. Dating speciation events with and without accounting for the among-site rate variation

Node (i)	Constant rate		Gamma rates	
	Distance (μt_i)	Time (t_i)	Distance (μt_i)	Time (t_i)
1	0.0627	15.77 ± 1.55	0.1024	18.91 ± 4.08
2	0.0517	13	0.0704	13
3	0.0307	7.73 ± 1.03	0.0362	6.68 ± 1.19
4	0.0201	5.04 ± 0.86	0.0239	4.41 ± 0.93
5	0.0093	2.35 ± 0.64	0.0100	1.85 ± 0.56

In a molecular clock analysis, the same substitution rate μ is assumed for all branches in the tree. The sequence data do not allow separation of time (t_i) from rate, and the distance (μt_i) is estimated instead, which measures the expected number of substitutions per site from the ancestral node i to the present time. A reference time obtained externally (for example, from fossil records) is used to calculate the substitution rate. In this example, the divergence time of orang-utan is fixed at 13 million years BP. Under the constant-rate-for-sites model, this leads to an evolutionary rate of $0.0517/(13\times10^6) = 3.9769\times10^{-9}$ substitutions per site per year. This rate is then used to convert other distances into times, for example, $t_4 = 0.0201/(3.9769\times10^{-9}) = 5.04\times10^6$ years BP for the separation of humans from chimpanzees. Under the gamma-rates model, the substitution rate is estimated to be $0.0704/(13\times10^6) = 5.4154\times10^{-9}$ substitutions per site per year, and the time of human–chimpanzee separation is estimated to be $t_4 = 4.41\times10^6$ years BP. As ignoring the rate variation tends to underestimate large distances more seriously than it underestimates small distances, the constant-rate model overestimates t_3, t_4, t_5 and underestimates t_1. The data are from 11 mitochondrial tRNA genes (759 nucleotides). *Adapted from Ref. 19.*



Effects of among-site rate variation

Estimation of evolutionary distances and speciation times

When evolutionary rates are variable among sites but are assumed to be constant, we tend to overlook some of the substitutions that have occurred at the fast-changing sites. Ignoring among-site rate variation therefore tends to give underestimates of sequence distances[5]. Gillespie[29] showed that when the Jukes–Cantor formula is used to calculate sequence distances when rates are gamma-distributed at sites, the calculated distance increases logarithmically (rather than linearly) with the true distance. Thus the formula consistently underestimates the true distance, with the bias being greater for large distances than for small ones.

The same pattern occurs with the estimation of branch lengths in a phylogenetic tree, branch length being measured by the average number of substitutions per site. For example, Box 2 shows a molecular clock analysis of mitochondrial tRNA genes to date speciation events among human, chimpanzee, gorilla, orang-utan and siamang, with the reference time (i.e. the divergence time of orang-utan) fixed at $t_2 = 13$ million years BP to 'calibrate' the molecular clock[19]. The estimated gamma parameter for these data is $\hat{\alpha} = 0.21 \pm 0.06$. The difference in log likelihood between the constant-rate model and the gamma-rates model, $\Delta\ell = 23.04$, is much greater than $\frac{1}{2}\chi^2 = 3.32$ at the 1% significance level, with one degree of freedom, indicating the existence of significant rate variation among sites. Because the lengths of long branches are more severely underestimated than those of short branches when rate variation among sites is ignored, the constant-rate model overestimates divergence times that are younger than the reference time (t_3, t_4, t_5), and underestimates divergence times that are older than the reference time (t_1). This effect was first reported by Adachi and Hasegawa[30], who used a discrete rate-class model to accommodate variable rates among sites.

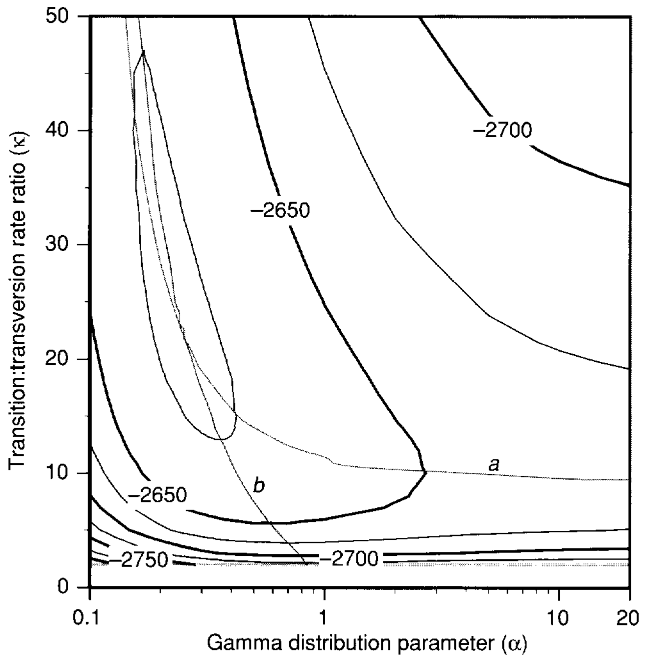


Fig. 2. The contour representation of the log-likelihood surface as a function of the gamma parameter α for rates at sites and the transition:transversion rate ratio (κ). The substitution model of Hasegawa *et al.*[14] is used and κ is the transition:transversion rate ratio (α/β in those authors' notation). Rates at sites are approximated by a discrete-gamma model[10]. Curve *a* represents the best estimate of κ when α is fixed at a specific value, while curve *b* represents the estimate of α when κ is fixed. The data are from a segment of mitochondrial DNA (895 bp) from human, chimpanzee, gorilla, orang-utan and gibbon[23]. The contour lines represent the log likelihood for given values of κ and α, maximized over the branch lengths.

Estimation of the transition rate bias

In a study of transition bias using pairwise sequence comparisons, Wakeley[31] noted that failure to account for the among-site rate variation leads to underestimation of the transition:transversion rate ratio. Intuitively, when transitions occur with higher frequency than transversions, many transitional substitutions are expected at the fast-changing sites. If the among-site rate variation is ignored, some of the transitions will be overlooked and the transition:transversion rate ratio will be underestimated[31,32].

A similar pattern was reported in maximum likelihood analyses[23,33]. Figure 2 shows a typical log-likelihood surface as a function of the gamma parameter, α, and the transition:transversion rate ratio, κ, highlighting the negative correlation between estimates of α and κ. The maximum likelihood estimates are $\hat{\alpha} = 0.25$ and $\hat{\kappa} = 21.98$. However, if rates are assumed to be constant among sites ($\alpha = \infty$), κ will be underestimated ($\hat{\kappa} = 9.39$), whereas if the transition:transversion rate bias is ignored (i.e. if κ is fixed at 1), α will be overestimated ($\hat{\alpha} = 0.77$) and the amount of rate variation underestimated. The correlation of estimates of the two parameters means that they should be estimated simultaneously. For the mitochondrial genome in which the transition bias is high, it is important to use an adequate substitution model[9,14,34] when estimating the gamma parameter.

Reconstruction of phylogenies

The existence of among-site rate variation means that most evolutionary changes occur at only a few sites, while many other sites never experience any substitutions. Since neither sites with very few evolutionary changes nor sites saturated with substitutions provide much phylogenetic information, sequences with severe rate variation tend to be less informative, even if the rate variation is adequately accounted for by the analytical method[27]. Furthermore, if the rate variation is present but ignored, model-based tree reconstruction methods, such as distance matrix methods and maximum likelihood methods, can be quite misleading, as shown by simulation studies and statistical consistency analysis[8,27,35-39]. The problem is especially acute if among-site rate variation is coupled with substitution rate variation among lineages. Simulations also show that the performance of the parsimony method in recovering the correct phylogeny deteriorates significantly when among-site rate variation exists[35,36,38]. Although the assumptions of parsimony are not explicitly specified, the method clearly involves some assumptions about rates as it performs worse when rates are variable among sites than when they are constant.

Furthermore, evaluation of the reliability of the estimated phylogeny appears to be quite sensitive to the assumed model[23,33]. Significance measures, such as the bootstrap proportions, are found to be dependent on whether the among-site rate variation is accounted for in the model, although the direction of the effect is not clear[33].

Conclusions and perspectives

Analyses of real data during the past few years have established that among-site rate variation exists and has important impact on various aspects of phylogenetic analysis, especially if the focus of the analysis is on the process of sequence evolution. It is, therefore, important to account for such rate variation in phylogenetic analysis. This can be done by adopting the maximum likelihood method, although the method involves intensive computation for large data sets. Distance matrix methods can also accommodate the rate variation if gamma distances with a reliable estimate of the α parameter are used. The current implementation of the parsimony method does not adequately deal with the among-site rate variation, although attempts have been made to allow for rate variation through unequal weighting of sites (characters)[40,41].

The past few years have also seen the development of reliable and practical likelihood methods for estimating the gamma parameter α. Methods using parsimony reconstructions of character changes to estimate α have also been improved, although even the improved method involves considerable positive bias. It may be possible to correct for the biases involved in the numbers of changes inferred by parsimony and to improve these estimation methods further. Results show that the reliability of the parsimony estimates increases with the number of sequences (taxa)[6,24]. Evolutionary biologists then have the option of using the likelihood method for small and medium-sized data sets and the improved parsimony methods for large data sets.

Simulation studies have also shown that the among-site rate variation has a significant influence on analyses of within-species data, such as the D-loop mitochondrial DNAs from human populations[42,43]. The rate variation shifts the distribution of the number of segregating sites in a DNA sample, invalidating Tajima's D statistic for testing neutrality[42]. It also causes the distribution of pairwise sequence differences to mimic patterns of population expansion. Much of the population genetics theory suitable for analyzing DNA sequence polymorphisms is developed under the infinite-sites model without accounting for among-site rate variation; the notion that these unrealistic assumptions do not matter due to the low divergence of within-species data appears to be a misconception. Appropriate analytical methods have yet to be developed.

Program availability

Most methods discussed in this review are implemented in the PAML program package (available at ftp.bio.indiana.edu:molbio/evolve). The discrete-gamma model[10] is also implemented in the PAUP* package. Calculation of pairwise distances under the gamma model is available in most phylogenetic packages including MEGA and PHYLIP.

Acknowledgements

I thank John Huelsenbeck, Masatoshi Nei, Sudhir Kumar, Rasmus Nelson, Bruce Rannala and Jack Sullivan for many constructive comments. Chris Simon and John Wakeley read an early version of this review in great detail, and their insightful comments are much appreciated. This work is partially supported by a grant from the National Science Foundation of China to Z.Y.

References

1 Fitch, W.M. and Margoliash, E. (1967) **A method for estimating the number of invariant amino acid coding positions in a gene using cytochrome *c* as a model case,** *Biochem. Genet.* 1, 65–71

2 Benzer, S. (1961) **On the topography of the genetic fine structure,** *Genetics* 47, 403–415

3 Li, W-H., Wu, C-I. and Luo, C-C. (1985) **Evolution of DNA sequences,** in *Molecular Evolutionary Genetics* (MacIntyre, J., ed.), pp. 1–94, Plenum Press

4 Uzzell, T. and Corbin, K.W. (1971) **Fitting discrete probability distributions to evolutionary events,** *Science* 172, 1089–1096

5 Golding, G.B. (1983) **Estimates of DNA and protein sequence divergence: an examination of some assumptions,** *Mol. Biol. Evol.* 1, 125–142

6 Wakeley, J. (1993) **Substitution rate variation among sites in hypervariable region 1 of human mitochondrial DNA,** *J. Mol. Evol.* 37, 613–623

7 Yang, Z. (1993) **Maximum likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites,** *Mol. Biol. Evol.* 10, 1396–1401

8 Jin, L. and Nei, N. (1990) **Limitations of the evolutionary parsimony method of phylogenetic analysis,** *Mol. Biol. Evol.* 7, 82–102

9 Tamura, K. and 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

10 Yang, Z. (1994) **Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods,** *J. Mol. Evol.* 39, 306–314

11 Rzhetsky, A. and Nei, M. (1994) **Unbiased estimates of the number of nucleotide substitutions when substitution rate varies among different sites,** *J. Mol. Evol.* 38, 295–299

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

13 King, J.L. and Jukes, T.H. (1969) **Non-darwinian evolution,** *Science* 164, 788–798

14 Hasegawa, M., Kishino, H. and Yano, T. (1985) **Dating the human–ape splitting by a molecular clock of mitochondrial DNA,** *J. Mol. Evol.* 22, 160–174

15 Palumbi, S.R. (1989) **Rates of molecular evolution and the function of nucleotide positions free to vary,** *J. Mol. Evol.* 29, 180–187

16 Gu, X., Fu, Y-X. and Li, W-H. (1995) **Maximum likelihood estimation of the heterogeneity of substitution rate among nucleotide sites,** *Mol. Biol. Evol.* 12, 546–557

17 Yang, Z. (1995) **A space-time process model for the evolution of DNA sequences,** *Genetics* 139, 993–1005

18 Hasegawa, M. *et al.* (1993) **Toward a more accurate time scale for the human mitochondrial DNA tree,** *J. Mol. Evol.* 37, 347–354

19 Yang, Z. **Maximum likelihood models for combined analyses of multiple sequence data,** *J. Mol. Evol.* (in press)

20 Felsenstein, J. and Churchill, G.A. (1996) **A hidden Markov model approach to variation among sites in rate of evolution,** *Mol. Biol. Evol.* 13, 93–104

21 Kelly, C. and Rice, J. (1996) **Modeling nucleotide evolution: a heterogeneous rate analysis,** *Math. Biosci.* 133, 85–109

22 Sullivan, J., Holsinger, K.E. and Simon, C. (1995) **Among-site rate variation and phylogenetic analysis of 12S rRNA in sigmontine rodents,** *Mol. Biol. Evol.* 12, 988–1001

23 Yang, Z., Goldman, N. and Friday, A.E. (1994) **Comparison of models for nucleotide substitution used in maximum likelihood phylogenetic estimation,** *Mol. Biol. Evol.* 11, 316–324

24 Sullivan, J., Holsinger, K.E. and Simon, C. (1996) **The effect of topology on estimation of among-site rate variation,** *J. Mol. Evol.* 42, 308–312

25 Yang, Z. and Wang, T. (1995) **Mixed model analysis of DNA sequence evolution,** *Biometrics* 51, 552–561

26 Kumar, S., Balczarek, K.A. and Lai, Z-C. (1996) **Evolution of the *hedgehog* gene family,** *Genetics* 142, 965–972

27 Yang, Z. (1995) **Evaluation of several methods for estimating phylogenetic trees when substitution rates differ over nucleotide sites,** *J. Mol. Evol.* 40, 689–697

28 Kumar, S. (1996) **Patterns of nucleotide substitution in mitochondrial protein coding genes of vertebrates,** *Genetics* 143, 537–548

29 Gillespie, J.H. (1986) **Rates of molecular evolution,** *Annu. Rev. Ecol. Syst.* 17, 637–665

30 Adachi, J. and Hasegawa, M. (1995) **Improved dating of the human/chimpanzee separation in the mitochondrial DNA tree: heterogeneity among amino acid sites,** *J. Mol. Evol.* 40, 622–628

31 Wakeley, J. (1994) **Substitution rate variation among sites and the estimation of transition bias,** *Mol. Biol. Evol.* 11, 436–442

32 Wakeley, J. (1996) **The excess of transitions among nucleotide substitutions: new methods of estimating transition bias underscore its significance,** *Trends Ecol. Evol.* 11, 158–163

33 Yang, Z., Goldman, N. and Friday, A.E. (1995) **Maximum likelihood trees from DNA sequences: a peculiar statistical estimation problem,** *Syst. Biol.* 44, 384–399

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

35 Tateno, Y., Takezaki, N. and Nei, M. (1994) **Relative efficiencies of the maximum-likelihood, neighbor-joining, and maximum-parsimony methods when substitution rate varies with site,** *Mol. Biol. Evol.* 11, 261–277

36 Kuhner, M.K. and Felsenstein, J. (1994) **A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates,** *Mol. Biol. Evol.* 11, 459–468

37 Gaut, B.S. and Lewis, P.O. (1995) **Success of maximum likelihood phylogeny inference in the four-taxon case,** *Mol. Biol. Evol.* 12, 152–162

38 Huelsenbeck, J.P. (1995) **The performance of phylogenetic methods in simulation,** *Syst. Biol.* 44, 17–48

39 Lockhart, P.J. *et al.* (1996) **Evolution of chlorophyll and bacteriochlorophyll: the problem of invariant sites in sequence analysis,** *Proc. Natl. Acad. Sci. U. S. A.* 93, 1930–1934

40 Farris, J.S. (1969) **A successive approximations approach to character weighting,** *Syst. Zool.* 18, 374–385

41 Williams, P.L. and Fitch, P.M. (1990) **Phylogenetic determination using dynamically weighted parsimony method,** *Methods Enzymol.* 183, 615–626

42 Berterolle, G. and Slatkin, M. (1995) **The number of segregating sites in expanding human populations, with implications for estimates of demographic parameters,** *Mol. Biol. Evol.* 12, 887–892

43 Aris-Brosou, S. and Excoffier, L. (1996) **The impact of population expansion and mutation rate heterogeneity on DNA sequence polymorphism,** *Mol. Biol. Evol.* 13, 494–504

44 Huelsenbeck, J.P., Cunningham, C.W. and Graybeal, A. **The performance of phylogenetic methods for a well-supported phylogeny,** *Syst. Biol.* (in press)

45 Rzhetsky, R. (1995) **Estimating substitution rates in ribosomal RNA genes,** *Genetics* 141, 771–783

46 Yang, Z., Lauder, I.J. and Lin, H.J. (1995) **Molecular evolution of the hepatitis B virus genome,** *J. Mol. Evol.* 41, 587–596

47 Jones, D.T., Taylor, W.R. and Thornton, J.M. (1992) **The rapid generation of mutation data matrices from protein sequences,** *Comput. Appl. Biosci.* 8, 275–282

Biodiversity and the productivity and stability of ecosystems

Kris H. Johnson, Kristiina A. Vogt, Heidi J. Clark, Oswald J. Schmitz and Daniel J. Vogt

Attempts to unveil the relationships between the taxonomic diversity, productivity and stability of ecosystems continue to generate inconclusive, contradictory and controversial conclusions. New insights from recent studies support the hypothesis that species diversity enhances productivity and stability in some ecosystems, but not in others. Appreciation is growing for the ways that particular ecosystem features, such as environmental variability and nutrient stress, can influence biotic interactions. Alternatives to the diversity–stability hypothesis have been proposed, and experimental approaches are starting to evolve to test these hypotheses and to elucidate the mechanisms underlying the functional role of species diversity.

Kris Johnson, Kristiina Vogt, Heidi Clark, Oswald Schmitz and Daniel Vogt are at the Yale School of Forestry and Environmental Studies, 205 Prospect Street, New Haven, CT 06511, USA (kjohnson@minerva.cis.yale.edu).

Resolution of the relationships between the diversity of life forms occupying ecosystems and the behavior of those systems is a prime directive of ecological research. The problem ultimately encompasses all questions about how species coexist and how communities of populations influence ecosystem performance. Early theoretical discussions established the axiom that diverse, complex ecological communities are the most stable[1]. Results of some field studies in the 1960s and 1970s began to challenge the universality of this paradigm. Theoretical advances based on mathematical modelling[1] indicated that the nature of species interactions, rather than species number alone, determines the stability of ecological systems. Subsequently, several hypotheses about diversity and ecosystem function relationships have been proposed. Recently, growing concern over the loss of biodiversity and new empirical evidence has prompted revisitation of the idea that species diversity enhances the productivity and stability of ecosystems[2].

Hypotheses and theoretical foundations

Four prevalent hypotheses (in addition to the null model) are summarized in Box 1: the diversity–stability, rivet or rivet-popper, redundancy, and idiosyncratic response hypotheses. The diversity–stability hypothesis[3] introduced the idea that increasing the number of trophically interacting species in an ecological community should increase the collective ability of member populations to maintain their abundances after disturbance. In his presentation of the hypothesis, MacArthur[3] implicitly recognized the transfer of energy from one trophic level to another as a quintessential ecosystem function. Biomass accumulation is the intuitively sensible, practical measure of energy assimilation, and field studies that followed used some measure of biomass as the ecosystem function response variable. The formalization of the diversity–stability hypothesis served as the original impetus for framing ecological questions in terms of the relationship between diversity and stability. The hypothesis came to be popularly acclaimed as a law of nature in spite of the fact that results of some field studies failed to support it[1].

The remaining hypotheses were introduced as alternatives to the diversity–stability hypothesis. The rivet hypothesis[4] suggests that ecosystem resistance – the ability of a system to absorb changes in abundances of some species without drastically changing ecosystem performance (e.g. biomass production)[5] – can decline as species are deleted even if system performance appears outwardly unaffected,

0169-5347/96/$15.00 PII: S0169-5347(96)10040-9