Statistical methods for detecting molecular adaptation

Ziheng Yang and Joseph P. Bielawski

The past few years have seen the development of powerful statistical methods for detecting adaptive molecular evolution. These methods compare synonymous and nonsynonymous substitution rates in protein-coding genes, and regard a nonsynonymous rate elevated above the synonymous rate as evidence for darwinian selection. Numerous cases of molecular adaptation are being identified in various systems from viruses to humans. Although previous analyses averaging rates over sites and time have little power, recent methods designed to detect positive selection at individual sites and lineages have been successful. Here, we summarize recent statistical methods for detecting molecular adaptation, and discuss their limitations and possible improvements.

Ziheng Yang and Joseph Bielawski are at the Galton Laboratory, Dept of Biology, University College London, 4 Stephenson Way, London, UK NW1 2HE (z.yang@ucl.ac.uk; j.bielawski@ucl.ac.uk).

'It has been proved remarkably difficult to get compelling evidence for changes in enzymes brought about by selection, not to speak of adaptive changes'[1].

Although Darwin's theory of evolution by natural selection is generally accepted by biologists for morphological traits (including behavioural and physiological), the importance of natural selection in molecular evolution has long been a matter of debate. The neutral theory[2] maintains that most observed molecular variation – both polymorphism within species and divergence between species – is due to random fixation of selectively neutral mutations. Well established cases of molecular adaptation have been rare[3]. Several tests of neutrality have been developed and applied to real data, and although they are powerful enough to reject strict neutrality in many genes, they rarely provide unequivocal evidence for positive darwinian selection.

Most convincing cases of adaptive molecular evolution have been identified through comparison of synonymous (silent; d_S) and nonsynonymous (amino acid-changing; d_N) substitution rates in protein-coding DNA sequences, thus providing fascinating case studies of natural selection in action on the protein molecule. Selected examples are listed in Table 1; see Hughes[4] for detailed descriptions of many case studies. Here, we summarize recent methodological developments that improve the power to detect adaptive molecular evolution, and examine their strengths and weaknesses , so that they can be used to detect more cases of molecular adaptation.

Measuring selection using the nonsynonymous/synonymous (d_N/d_S) rate ratio

Traditionally, synonymous and nonsynonymous substitution rates (Box 1) are defined in the context of comparing two DNA sequences, with d_S and d_N as the numbers of synonymous and nonsynonymous substitutions per site, respectively[5]. Thus, the ratio $\omega = d_N/d_S$ measures the difference between the two rates and is most easily understood from a mathematical description of a codon substitution model (Box 2). If an amino acid change is neutral, it will be fixed at the same rate as a synonymous mutation, with $\omega = 1$. If the amino acid change is deleterious, purifying selection (Box 1) will reduce its fixation rate, thus $\omega < 1$. Only when the amino acid change offers a selective advantage is it fixed at a higher rate than a synonymous mutation, with $\omega > 1$. Therefore, an ω ratio significantly higher than one is convincing evidence for diversifying selection.

The codon-based analysis (Box 2) cannot infer whether synonymous substitutions are driven by mutation or selection, but it does not assume that synonymous substitutions are neutral. For example, highly biased codon usage can be caused by both mutational bias and selection (e.g. for translational efficiency[6]), and can greatly affect synonymous substitution rates. However, by employing parameters π_j for the frequency of codon j in the model (Box 2), estimation of

 0169-5347/00/$ – see front matter PII: S0169-5347(00)01994-7

Table 1. Selected examples of protein-coding genes in which positive selection was detected by using the d_N/d_S ratio

Gene	Organism	Refs
Genes involved in defensive systems or immunity		
Class I chitinase gene	*Arabis* and *Arabidopsis*	41
Colicin genes	*Escherichia coli*	45
Defensin genes	Rodents	46
Fv1	*Mus*	47
Immunoglobulin V_H genes	Mammals	48
MHC genes	Mammals	49
Polygalacturonase inhibitor genes	Legume and dicots	50
RH blood group and RH50 genes	Primates and rodents	51
Ribonuclease genes	Primates	52
Transferrin gene	Salmonid fishes	53
Type I interferon-ω gene	Mammals	54
α_1-Proteinase inhibitor genes	Rodents	55
Genes involved in evading defensive systems or immunity		
Capsid gene	FMD virus	42
CSP, TRAP, MSA-2 and *PF83*	*Plasmodium falciparum*	56
Delta-antigen coding region	Hepatitis D virus	57
E gene	Phages *G4*, *ϕX174*, and *S13*	3
Envelope gene	HIV	40
gH glycoprotein gene	Pseudorabies virus	3
Hemagglutinin gene	Human influenza A virus	33
Invasion plasmid antigen genes	*Shigella*	3
Merozoite surface antigen-1 gene	*Plasmodium falciparum*	58
*msp 1*α	*Anaplasma marginale*	3
nef	HIV	38
Outer membrane protein gene	*Chlamydia*	3
Polygalacturonase genes	Fungal pathogens	50
Porin protein 1 gene	*Neisseria*	59
S and *HE* glycoprotein genes	Murine coronavirus	60
Sigma-1 protein gene	Reovirus	3
Virulence determinant gene	*Yersinia*	3
Genes involved in reproduction		
18-kDa fertilization protein gene	Abalone (*Haliotis)*	61
Acp26Aa	*Drosophila*	62
Androgen-binding protein gene	Rodents	63
Bindin gene	*Echinometra*	64
Egg-laying hormone genes	*Aplysia californica*	3
Ods homeobox gene	*Drosophila*	65
Pem homeobox gene	Rodents	66
Protamine P1 gene	Primates	67
Sperm lysin gene	Abalone (*Haliotis)*	61
S-Rnase gene	Rosaceae	68
Sry gene	Primates	69
Genes involved in digestion		
κ-casein gene	Bovids	70
Lysozyme gene	Primates	23
Toxin protein genes		
Conotoxin genes	*Conus* gastropods	71
Phospholipase A_2 gene	Crotalinae snakes	72
Genes related to electron transport and/or ATP synthesis		
ATP synthase F_0 subunit gene	*Escherichia coli*	3
COX7A isoform genes	Primates	73
COX4 gene	Primates	74
Cytokine genes		
Granulocyte-macrophage SF gene	Rodents	75
Interleukin-3 gene	Primates	75
Interleukin-4 gene	Rodents	75
Miscellaneous		
CDC6	*Saccharomyces cerevisiae*	3
Growth hormone gene	Vertebrates	76
Hemoglobin β-chain gene	Antarctic fishes	77
Jingwei	*Drosophila*	78
Prostatein peptide C3 gene	Rat	3

substitution rates will fully account for codon-usage bias (Box 1), irrespective of its source. Because parameter ω is a measure of selective pressure on a protein, it differentiates codon-based analyses from the more general tests of neutrality proposed in population genetics[7,8]. These general tests often lack the power to determine the sources of the departure from the strict neutral model, such as changes in population size, fluctuating environment or different forms of selection.

Estimation of d_N and d_S between two sequences

Two classes of methods have been suggested to estimate d_N and d_S between two protein-coding DNA sequences. The first class includes over a dozen intuitive methods developed since the early 1980s (Refs 5,9–15). These methods involve the following steps: counting synonymous (*S*) and nonsynonymous (*N*) sites in the two sequences, counting synonymous and nonsynonymous differences between the two sequences, and correcting for multiple substitutions at the same site. *S* and *N* are defined as the sequence length multiplied by the proportions of synonymous and nonsynonymous changes before selection on the protein[14,16]. Most of these methods make simplistic assumptions about the nucleotide substitution process and also involve *ad hoc* treatment of the data that cannot be justified[14,15]; therefore, we refer to these methods of estimating d_N and d_S as approximate methods. The methods of Miyata and Yasunaga[5], and Nei and Gojobori[9], assume an equal rate for transitions (T ↔ C and A ↔ G) and transversions (T,C ↔ A,G), as well as a uniform codon usage. Because transitions at the third 'wobble' position are more likely to be synonymous than transversions, ignoring the transition/transversion rate ratio leads to underestimation of *S* and overestimation of *N* (Ref. 10). Efforts have been taken to incorporate the transition/transversion rate bias (Box 1) when counting sites and differences[10–14]. The effect of

Box 1. Glossary

Codon-usage bias: unequal codon frequencies in a gene.
Nonsynonymous substitution: a nucleotide substitution that changes the encoded amino acid.
Prior probability: the probability of an event (such as a site belonging to a site class) before the collection of data.
Positive selection: darwinian selection fixing advantageous mutations with positive selective coefficients. The term is used interchangeably with molecular adaptation and adaptive molecular evolution.
Posterior probability: the probability of an event conditional on the observed data, which reflects both the prior assumption and information in the data.
Purifying selection: natural selection against deleterious mutations with negative selective coefficients. The term is used interchangeably with negative selection or selective constraints.
Synonymous substitution: a nucleotide substitution that does not change the encoded amino acid.
Transition/transversion rate bias: unequal substitution rates between nucleotides, with a higher rate for transitions (changes between T and C and between A and G) than transversions (all other changes).

Box 2. A model of codon substitution

The codon is considered the unit of evolution. The substitution rate from codons i to j ($i \neq j$) is given as:

$$q_{ij} = \begin{cases} 0, & \text{if } i \text{ and } j \text{ differ at more than one position,} \\ \pi_j, & \text{for synonymous transversion,} \\ \kappa\pi_j, & \text{for synonymous transition,} \\ \omega\pi_j, & \text{for nonsynonymous transversion,} \\ \omega\kappa\pi_j, & \text{for nonsynonymous transition.} \end{cases}$$

Parameter κ is the transition/transversion rate ratio, π_j is the equilibrium frequency of codon j and ω ($= d_N/d_S$) measures the selective pressure on the protein. The q_{ij} are relative rates because time and rate are confounded in such an analysis. Given the rate matrix $Q = \{q_{ij}\}$, the transition probability matrix over time t is calculated as:

$$P(t) = \{p_{ij}(t)\} = e^{Qt}$$

where $p_{ij}(t)$ is the probability that codon i becomes codon j after time t. Likelihood calculation on a phylogeny involves summing over all possible codons in extinct ancestors (internal nodes of the tree). After Refs 16,18,27,79.

biased codon usage has largely been ignored[17]; however, extreme codon-usage bias can have devastating effects on the estimation of d_N and d_S (see the next section)[15,18]. A recent *ad hoc* method[15] incorporates both transition and codon-usage biases.

The second class is the maximum likelihood (ML) method based on explicit models of codon substitution (Box 2)[16,19]. Parameters in the model (i.e. sequence divergence t, transition/transversion rate ratio κ and the d_N/d_S ratio ω) are estimated from the data by ML, and are used to calculate d_N and d_S according to their definitions[15,16,20]. A major feature of the method is that the model is formulated at the level of instantaneous rates (where there is no possibility for multiple changes) and that probability theory accomplishes all difficult tasks in one step: estimating mutational parameters, such as κ; correcting for multiple hits; and weighting pathways of change between codons.

Statistical tests can be used to test whether d_N is significantly higher than d_S. For approximate methods, a normal approximation is applied to $d_N - d_S$. For ML, a likelihood-ratio test can be used. In this case, the null model has ω fixed at 1, whereas the alternative model estimates ω as a free parameter. Twice the log-likelihood difference between the two models is compared with a χ^2 distribution with one degree of freedom to test whether ω is different from one.

Computer simulation has been used to examine the performance of different estimation methods; the findings are consistent with observations made in real data analyses[14,15,19]. We demonstrate the effects of different estimation procedures using human and orangutan α_2-globin genes (Table 2). For comparison, different assumptions are made in ML concerning the transition/transversion rate bias and the codon-usage bias. The simpler models are each rejected when compared with more complex models by likelihood-ratio tests, confirming biased transition rates and codon usage. Thus, estimates from ML accounting for both biases (Model 8, Table 2) are expected to be the most reliable. We make the following observations:

- Assumptions appear to matter more than methods. The approximate methods and ML produce similar results under similar assumptions. The method of Nei and Gojobori is similar to ML under a model that ignores both transition/transversion bias and codon-usage bias (Model 1, Table 2), whereas the methods of Ina and Li are similar to ML under a model accounting for the transition/transversion bias but ignoring codon-usage bias (Model 2, Table 2). The method of Yang and Nielsen[15] is similar to ML under a model accounting for both biases (Model 6, Table 2). However, for distantly related sequences, *ad hoc* treatment in approximate methods can lead to serious biases even under the correct assumptions[19].
- Ignoring the transition/transversion rate bias leads to underestimation of S, overestimation of d_S and underestimation of the ω ratio[10].
- Codon-usage bias in these data has the opposite

Table 2. Estimation of d_N and d_S between the human and orangutan α_2-globin genes (142 codons)[a]

Method and/or model	κ	S	N	d_N	d_S	d_N/d_S (ω)	ℓ[c]	Refs
Approximate methods								
Nei and Gojobori	1.0	109.4	316.6	0.0095	0.0569	0.168	–	9
Li	–	NA	NA	0.0104	0.0517	0.201	–	11
Ina	2.1	119.3	299.9	0.0101	0.0523	0.193	–	14
Yang and Nielsen	6.1	61.7	367.3	0.0083	0.1065	0.078	–	15
ML methods[b]								
(1) Fequal, $\kappa = 1$	1.0	108.5	317.5	0.0093	0.0557	0.167	−633.67	16
(2) Fequal, κ estimated	3.0	124.6	301.4	0.0099	0.0480	0.206	−632.47	16
(3) F1×4, $\kappa = 1$ fixed	1.0	129.1	296.9	0.0092	0.0671	0.137	−612.40	16
(4) F1×4, κ estimated	3.9	137.1	288.9	0.0093	0.0624	0.149	−610.48	16
(5) F3×4, $\kappa = 1$ fixed	1.0	63.2	362.8	0.0084	0.0973	0.087	−560.76	16
(6) F3×4, κ estimated	5.4	60.6	365.4	0.0084	0.1061	0.079	−557.85	16
(7) F61, $\kappa = 1$ fixed	1.0	58.3	367.7	0.0082	0.1145	0.072	−501.39	16
(8) F61, κ estimated	5.3	55.3	370.7	0.0082	0.1237	0.066	−498.61	16

[a]GenBank accession numbers are V00516 (human) and M12158 (orangutan).
[b]Fequal, equal codon frequencies (= 1/61) are assumed; F1×4, four nucleotide frequencies are used to calculate codon frequencies (3 free parameters); F3×4, nucleotide frequencies at three codon positions are used to calculate codon frequencies (9 free parameters); F61, all codon frequencies are used as free parameters (60 free parameters).
[c]ℓ is the log-likelihood value.

effect to the transition/transversion bias; ignoring codon-usage bias leads to overestimation of S, underestimation of d_S and overestimation of ω. This gene is extremely GC-rich at the third codon position, with base frequencies at 9% (T), 52% (C), 1% (A) and 37% (G). Most changes at the third position (before selection at the amino acid level) are transversions between C and G. Thus, the number of synonymous sites is less than half that expected under equal base and codon frequencies. Although, in theory, the bias caused by unequal codon frequencies can be in the opposite direction[15], we have not encountered a real gene showing that pattern. Such codon-usage bias appears to have misled previous analyses examining the relationship between the GC content at silent sites and d_S, because those studies ignored the codon-usage bias when estimating d_S (Ref. 21).

- Different methods can produce different estimates, even when the sequences are highly similar. The sequences used in Table 2 are only about 10% different at silent sites and <1% different at nonsynonymous sites; however, estimates of ω are three times different. Because all estimation procedures partition the total numbers of sites and differences into synonymous and nonsynonymous categories, underestimation of one means overestimation of the other, thus resulting in large errors in the ω ratio.

Detecting lineage-specific episodes of darwinian selection

If, for most of the time, a gene evolves under purifying selection but is occasionally subject to episodes of adaptive change[22], a comparison between two distantly related sequences is unlikely to yield a d_N/d_S ratio significantly greater than one. Methods have been developed to detect positive selection (Box 1) along specific lineages on a phylogeny. If the gene sequences of the extinct ancestors were known, it would be straightforward to use the pairwise methods discussed above. Thus, Messier and Stewart[23] inferred ancestral lysozyme gene sequences through phylogenetic analysis[24,25], and used them to calculate d_N and d_S for each branch in the phylogeny. Their analysis identified two lineages in a primate phylogeny with highly elevated nonsynonymous substitution rates. The same approach was taken in a test of relaxed selective constraint in the rhodopsin gene of cave-dwelling crayfishes[26].

There are also likelihood models that allow different ω ratios for branches in a phylogeny[18,27]. Using such models, likelihood-ratio tests can be constructed to test hypotheses. For example, the ω ratio for a predefined lineage can be either fixed at one or estimated as a free parameter. The likelihood values under those two models can be compared, to test whether $\omega > 1$ in that lineage. Similarly, a model assuming a single ω for all lineages (the one-ratio model) can be compared with another model assuming an independent ω for each lineage (the free-ratio model), to test the neutral prediction that the ω ratio is identical among lineages[18,27].

It should be noted that variation in the ω ratio among lineages is a violation of the strictly neutral model[2,18,28,29], but it is not sufficient evidence for adaptive evolution. In particular, if nonsynonymous mutations are slightly deleterious, they will have a higher probability of fixation in a small population than in a large one[30], and thus lineages of different population sizes will have different ω ratios. Besides positive selection, relaxed selective constraint can also elevate the ω ratio – it might be difficult to distinguish the two if the estimated ω is not larger than one. Furthermore, it is incorrect to use the free-ratio model to identify lineages of interest and then to perform further tests on the ω ratios for those lineages using the same data without any correction[27].

Box 3. Likelihood and Bayes

The statistical-estimation theory used in the methods discussed in this review can be explained with the following simple hypothetical example. Suppose that a population is an admixture of two groups of people in the proportions 60% and 40%, and a certain disease occurs at a rate of 1% in Group I and of 0.1% in Group II. Suppose a random sample of 100 individuals is taken from the population, what is the probability that three of them carry the disease? The probability that a random individual carries the disease (D) is an average over the two groups (G_1 and G_2):

$$p = P(D) = P(G_1) \times P(D \mid G_1) + P(G_2) \times P(D \mid G_2) = 0.6 \times 0.01 + 0.4 \times 0.001 = 0.0064 \quad (1)$$

Similarly, the probability that an individual does not carry the disease is:

$$P(\bar{D}) = P(G_1) \times P(\bar{D} \mid G_1) + P(G_2) \times P(\bar{D} \mid G_2)$$
$$= 0.6 \times 0.99 + 0.4 \times 0.999 = 0.9936 = 1 - p \quad (2)$$

The probability that three out of 100 individuals carry the disease is given by the binomial probability:

$$P = \frac{100!}{3!\times 97!}\left[p^3(1-p)^{97}\right] = 0.0227 \quad (3)$$

If Eqn 3 involves an unknown parameter [such as the rate $P(D|G_1)$ in Group I], that parameter can be estimated by maximizing Eqn 3. In that case, Eqn 3 gives the probability of observing the data (sample) and is called the likelihood function.

The second question is to calculate the probability that an individual in the sample who carries the disease is from Group I. The Bayes theorem gives this probability as:

$$P(G_1 \mid D) = P(G_1) \times P(D \mid G_1)/P(D) = 0.6 \times 0.01/0.0064 = 0.94 \quad (4)$$

Note that this is just the proportion of the contribution from Group I to $P(D)$ in Eqn 1. Thus, this individual is most likely to be from Group I. Similarly, a healthy individual in the sample is more likely to be from Group II than from Group I because

$$P(G_1 \mid \bar{D}) = P(G_1) \times P(\bar{D} \mid G_1)/P(\bar{D}) = 0.6 \times 0.99/0.9936 = 0.5978$$
$$\text{and } P(G_2 \mid \bar{D}) = 1 - P(G_1 \mid \bar{D}) = 0.4022 \quad (5)$$

In methods for inferring sites under positive selection[36,37], we let D in the example be the data at a site and G_i be the ith site class with the d_N/d_S ratio ω_i. The probability of observing data at a site is then an average over the site classes (Eqn 1). The product of such probabilities over sites constitutes the likelihood (Eqn 3), from which we estimate any unknown parameters, such as the branch lengths and parameters in the ω distribution over sites. After the parameters are estimated, we use the Bayes theorem to calculate the probability that any site, given data at that site, is from each site class (Eqns 4 and 5).

Another straightforward application of the theory is ancestral sequence reconstruction; in this case, we replace G_i with a reconstruction (characters at interior nodes of the phylogeny) at a site. When we calculate the likelihood function, the probability of data at a site $P(D)$ is a sum over all possible ancestral reconstructions (G_is) (Eqns 1 and 2). After parameters are estimated, the reconstruction that makes the greatest contribution to $P(D)$ is the most likely (Eqns 4 and 5)[24].

The Bayes method discussed here is known as the empirical Bayes, because it uses estimates of parameters and does not account for their sam-pling errors. This might be a concern if parameters are estimated from small samples or if the posterior probabilities are sensitive to parameter estimates. An alternative approach is the hierarchical Bayes method, which accounts for the uncertainty in unknown parameters by averaging over their prior distribution.

Note that the reconstructed ancestral sequences[24], as well as the inferred site classes in the site-class models[36,37], are pseudo data and involve systematic biases. To appreciate such biases, note that in the previous example, the Bayes calculations (Eqns 4 and 5) predict that each of the 100 individuals in the sample, healthy or sick, are from Group I. Although this is the best prediction, the accuracy is low. If such inferred group identities are used for further statistical analysis, misleading results might follow.

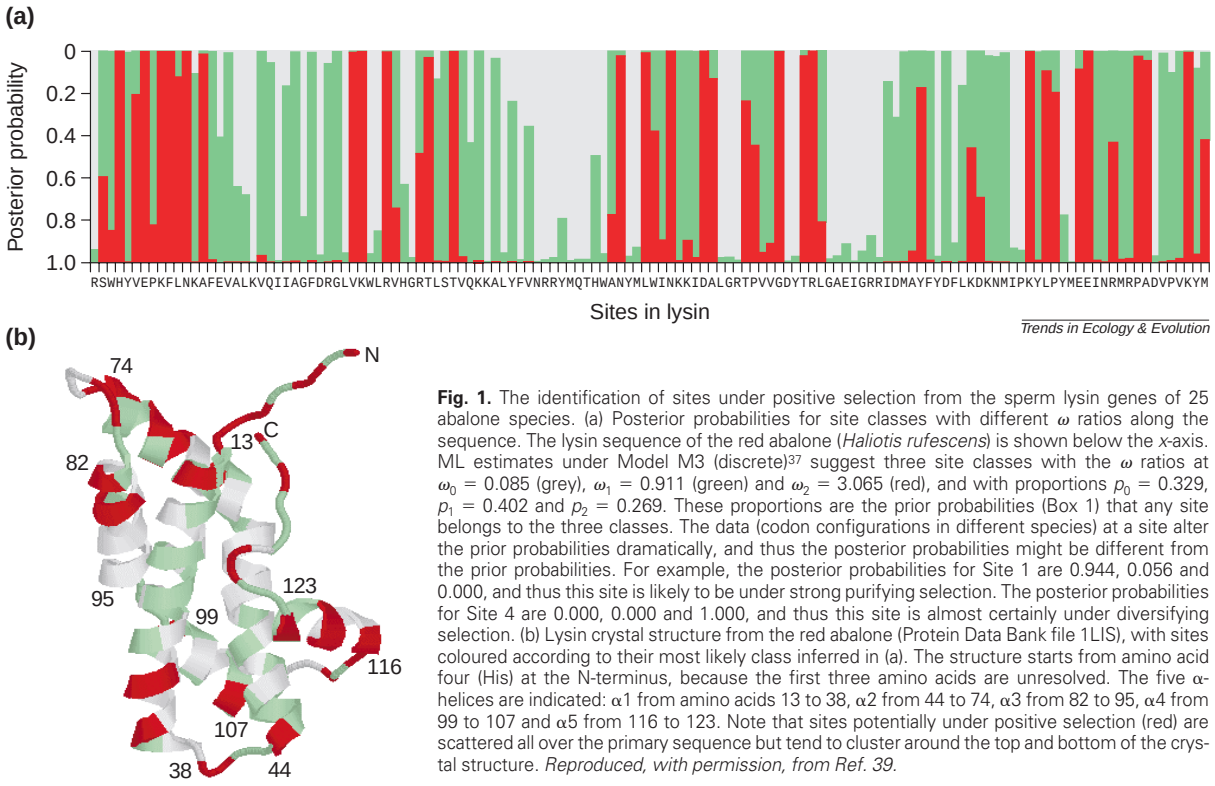

Fig. 1. The identification of sites under positive selection from the sperm lysin genes of 25 abalone species. (a) Posterior probabilities for site classes with different ω ratios along the sequence. The lysin sequence of the red abalone (*Haliotis rufescens*) is shown below the *x*-axis. ML estimates under Model M3 (discrete)[37] suggest three site classes with the ω ratios at $\omega_0 = 0.085$ (grey), $\omega_1 = 0.911$ (green) and $\omega_2 = 3.065$ (red), and with proportions $p_0 = 0.329$, $p_1 = 0.402$ and $p_2 = 0.269$. These proportions are the prior probabilities (Box 1) that any site belongs to the three classes. The data (codon configurations in different species) at a site alter the prior probabilities dramatically, and thus the posterior probabilities might be different from the prior probabilities. For example, the posterior probabilities for Site 1 are 0.944, 0.056 and 0.000, and thus this site is likely to be under strong purifying selection. The posterior probabilities for Site 4 are 0.000, 0.000 and 1.000, and thus this site is almost certainly under diversifying selection. (b) Lysin crystal structure from the red abalone (Protein Data Bank file 1LIS), with sites coloured according to their most likely class inferred in (a). The structure starts from amino acid four (His) at the N-terminus, because the first three amino acids are unresolved. The five α-helices are indicated: α1 from amino acids 13 to 38, α2 from 44 to 74, α3 from 82 to 95, α4 from 99 to 107 and α5 from 116 to 123. Note that sites potentially under positive selection (red) are scattered all over the primary sequence but tend to cluster around the top and bottom of the crystal structure. *Reproduced, with permission, from Ref. 39.*

Methods based on ancestral reconstruction might not provide reliable statistical tests because they ignore errors and biases in reconstructed ancestral sequences (Box 3). The ML method has the advantage of not relying on reconstructed ancestral sequences. It can also easily incorporate features of DNA sequence evolution, such as the transition/transversion rate bias and codon-usage bias, and is thus based on a more realistic evolutionary model. When likelihood-ratio tests suggest adaptive evolution along certain lineages, ancestral reconstruction might be useful to pinpoint the involved amino acids and to infer ancestral proteins, which can be synthesized and examined in the laboratory[31,32].

Detecting amino acid sites under darwinian selection

The methods discussed so far assume that all amino acid sites are under the same selective pressure, with the same ω ratio. The analysis effectively averages the ω ratio across all sites and positive selection is detected only if that average is >1. This appears to be a conservative test of positive selection because many sites might be under strong purifying selection owing to functional constraint, with the ω ratio close to zero.

A few recent studies addressed this problem. Fitch and colleagues[33,34] used parsimony to reconstruct ancestral DNA sequences, and counted changes at each codon site along branches of the tree. They tested whether the proportion of nonsynonymous substitutions at each site is greater than the average over all sites in the sequence. Suzuki and Gojobori[35] took a more systematic approach. For each site in the sequence, they estimated the numbers of synonymous and nonsynonymous sites and differences along the tree using reconstructed ancestral sequences, and then tested whether the proportion of nonsynonymous substitutions differed from the neutral expectation ($\omega = 1$). Suzuki and Gojobori's criterion is more stringent than Fitch *et al.*'s, because the ω ratio averaged over sites is almost always <1. These methods are expected to require many sequences in the data set so that there are enough changes at individual sites. Furthermore, the reliability of significance values produced by these methods might be affected by the use of ancestral reconstruction, which is most unreliable at the positively selected or variable sites[24], and by codon composition bias, which is most extreme at a single site.

In a likelihood model, it is impractical to use one ω parameter for each site. The standard approach is to use a statistical distribution to describe the variation of ω among sites; for example, we might assume several classes of sites in the protein with different ω ratios[36,37]. The test of positive selection then involves two major steps: first, to test whether sites exist where $\omega > 1$, which is achieved by a likelihood-ratio test comparing a model that does not allow for such sites with a more general model that does; and second, to use the Bayes theorem to identify positively selected sites when they exist. Sites having high posterior probabilities (Box 1) for site classes with $\omega > 1$ are potential targets of diversifying selection. The theory is explained in Box 3 (Refs 20,36,37).

Nielsen and Yang[36] implemented a likelihood-ratio test based on two simple models. The null model, M1 (neutral), assumes a class of conserved sites with $\omega = 0$ and another class of neutral sites with $\omega = 1$. The alternative model, M2 (selection), adds a third class of sites with ω estimated from the data. (The model codes are those used in the codeml program in the PAML package.) If M2 fits the data significantly better than M1 and the estimated ω ratio for the third class in M2 is >1, then some sites are under

diversifying selection. Zanotto *et al.*[38] used this test to identify several sites under strong positive selection in the *nef* gene of HIV, whereas both pairwise comparison and sliding-window analysis failed. This comparison was later found to lack power in some genes because M1 does not account for sites with $0 < \omega < 1$ and the third class in M2 is forced to account for such sites[37]. Thus, Yang *et al.*[37] implemented several new models. For example, the beta distribution (M7 beta) is a flexible null model with $0 < \omega < 1$, and can be compared with an alternative that adds an additional site class with ω estimated (M8 beta&ω). A general discrete model (M3) was also implemented[37]. These models identified positive selection in six out of ten genes the authors analysed. Figure 1 shows the use of a discrete model (M3) with three classes to identify sites under diversifying selection in abalone sperm lysin[39].

The methods discussed above assume that there are heterogeneous classes of amino acid sites but that we do not know *a priori* which class each site is from. Such 'fishing-expedition' studies might be useful in generating hypotheses for laboratory investigation because they could identify crucial amino acids whose changes have offered a selective advantage in Nature's evolutionary experiment. For example, amino acid residues under diversifying selection were inferred in analyses of HIV-1 *nef* (Ref. 38) and *env* (Ref. 40) genes, which might constitute unidentified viral epitopes. Alternatively, we might wish to test an *a priori* hypothesis that certain structural and functional domains of the protein are under positive selection. In such cases, likelihood models can be constructed that assign and estimate different ω parameters for sites from different structural and functional domains[20].

Limitations of current methods and future directions

All the methods for detecting positive selection reviewed here appear to be conservative. They detect selection only if d_N is higher than d_S – selection that does not cause excessive nonsynonymous substitutions, such as balancing selection, might not be detected. The pairwise comparison has little power because it averages the ω ratio over sites and over time. Methods for detecting selection along lineages work only if the ω ratio averaged over all sites is >1. Similarly, the test of positive selection at sites works only if the ω ratio averaged over all branches is >1. If adaptive evolution occurs only in a short time interval and affects only a few crucial amino acids, none of the methods is likely to succeed. Constancy of selective pressure at sites appears to be a much more serious assumption than constancy among lineages, especially for genes likely to be under continuous selective pressure, such as the HIV *env* gene. Indeed, models of variable selective pressures among sites[36,37] have been successful in detecting positive selection, even in a background of overwhelming purifying selection indicated by an average ω ratio much smaller than one[37,38,41,42]. Models that allow ω to vary among both lineages and sites should have increased power.

The methods discussed here also assume the same ω ratio for all possible amino acid changes; for example, at a positively selected site, all amino acid changes are assumed to be advantageous, which is unrealistic. Although amino acid substitution rates are known to correlate with their chemical properties, the relationship is poorly understood[43,44]. It is also not entirely clear how to define positive selection in a model accounting for chemical properties.

It will be interesting to perform computer simulations to examine the power of various detection methods and to investigate how this is affected by important factors, such as the size of the gene, sampling of species (sequences) and the level of sequence divergence. Including more sequences in the data should improve the power of site-based analyses. Sequence divergence is also important because neither very similar nor very divergent sequences contain much information. Very divergent sequences might also be associated with problems with alignment and unequal nucleotide compositions in different species. Analyses discussed here, which require information from both synonymous and nonsynonymous substitutions, are expected to have a narrower window of suitable sequence divergences than phylogeny reconstruction. The large-sample χ^2 approximation to the likelihood-ratio test statistic might also be examined, but limited simulations suggest that typical sequence data (with >100 codons) are large enough for it to be reliable. For very short genes or gene regions and especially at low sequence divergences, Monte Carlo simulation might be needed to derive the null distribution.

The likelihood analysis assumes no recombination within a gene. If recombination occurs, different regions will have different phylogenies. Empirical data analysis suggests that the phylogeny does not have much impact on tests of positive selection and identification of sites, and one might suspect that recombination will not cause false positives by the likelihood-ratio test. However, simulation studies are necessary to understand whether this is the case.

Acknowledgements

We thank D. Haydon, J. Mallet, T. Ohta, A. Pomiankowski, V. Vacquier, W. Swanson and three anonymous referees for comments. We also thank several users of the PAML package (http://abacus.gene.ucl.ac.uk/software/paml.html), in particular C. Woelk, for comments and suggestions concerning the implementation. This work is supported by grant #31/G10434 from the Biotechnology and Biological Sciences Research Council (UK).

References

1 Lewontin, R.C. (1979) Adaptation. *Sci. Am.* 239, 156–169
2 Kimura, M. (1983) *The Neutral Theory of Molecular Evolution*, Cambridge University Press
3 Endo, T. *et al.* (1996) Large-scale search for genes on which positive selection may operate. *Mol. Biol. Evol.* 13, 685–690
4 Hughes, A.L. (1999) *Adaptive Evolution of Genes and Genomes*, Oxford University Press
5 Miyata, T. and Yasunaga, T. (1980) Molecular evolution of mRNA: a method for estimating evolutionary rates of synonymous and amino acid substitutions from homologous nucleotide sequences and its applications. *J. Mol. Evol.* 16, 23–36
6 Akashi, H. (1995) Inferring weak selection from patterns of polymorphism and divergence at 'silent' sites in *Drosophila* DNA. *Genetics* 139, 1067–1076
7 Kreitman, M. and Akashi, H. (1995) Molecular evidence for natural selection. *Annu. Rev. Ecol. Syst.* 26, 403–422
8 Wayne, M.L. and Simonsen, K.L. (1998) Statistical tests of neutrality in the age of weak selection. *Trends Ecol. Evol.* 13, 236–240
9 Nei, M. and Gojobori, T. (1986) Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol. Biol. Evol.* 3, 418–426
10 Li, W-H. *et al.* (1985) A new method for estimating synonymous and nonsynonymous rates of nucleotide substitutions considering the relative likelihood of nucleotide and codon changes. *Mol. Biol. Evol.* 2, 150–174
11 Li, W-H. (1993) Unbiased estimation of the rates of synonymous and nonsynonymous substitution. *J. Mol. Evol.* 36, 96–99
12 Pamilo, P. and Bianchi, N.O. (1993) Evolution of the *Zfx* and *Zfy* genes – rates and interdependence between the genes. *Mol. Biol. Evol.* 10, 271–281

13 Comeron, J.M. (1995) A method for estimating the numbers of synonymous and nonsynonymous substitutions per site. *J. Mol. Evol.* 41, 1152–1159

14 Ina, Y. (1995) 0New methods for estimating the numbers of synonymous and nonsynonymous substitutions. *J. Mol. Evol.* 40, 190–226

15 Yang, Z. and Nielsen, R. (2000) Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Mol. Biol. Evol.* 17, 32–43

16 Goldman, N. and Yang, Z. (1994) A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol. Biol. Evol.* 11, 725–736

17 Moriyama, E.N. and Powell, J.R. (1997) Synonymous substitution rates in *Drosophila*: mitochondrial versus nuclear genes. *J. Mol. Evol.* 45, 378–391

18 Yang, Z. and Nielsen, R. (1998) Synonymous and nonsynonymous rate variation in nuclear genes of mammals. *J. Mol. Evol.* 46, 409–418

19 Muse, S.V. (1996) Estimating synonymous and nonsynonymous substitution rates. *Mol. Biol. Evol.* 13, 105–114

20 Yang, Z. (2000) Adaptive molecular evolution. In *Handbook of Statistical Genetics* (Balding, D. *et al.*, eds), Ch. 12, Wiley

21 Bielawski, J. *et al.* Rates of nucleotide substitution and mammalian nuclear gene evolution: approximate and maximum-likelihood methods lead to different conclusions. *Genetics* (in press)

22 Gillespie, J.H. (1991) *The Causes of Molecular Evolution*, Oxford University Press

23 Messier, W. and Stewart, C-B. (1997) Episodic adaptive evolution of primate lysozymes. *Nature* 385, 151–154

24 Yang, Z. *et al.* (1995) A new method of inference of ancestral nucleotide and amino acid sequences. *Genetics* 141, 1641–1650

25 Koshi, J.M. and Goldstein, R.A. (1996) Probabilistic reconstruction of ancestral protein sequences. *J. Mol. Evol.* 42, 313–320

26 Crandall, K.A. and Hillis, D.M. (1997) Rhodopsin evolution in the dark. *Nature* 387, 667–668

27 Yang, Z. (1998) Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. *Mol. Biol. Evol.* 15, 568–573

28 McDonald, J.H. and Kreitman, M. (1991) Adaptive protein evolution at the *Adh* locus in *Drosophila*. *Nature* 351, 652–654

29 Hasegawa, M. *et al.* (1998) Preponderance of slightly deleterious polymorphism in mitochondrial DNA: replacement/synonymous rate ratio is much higher within species than between species. *Mol. Biol. Evol.* 15, 1499–1505

30 Ohta, T. (1973) Slightly deleterious mutant substitutions in evolution. *Nature* 246, 96–98

31 Golding, G.B. and Dean, A.M. (1998) The structural basis of molecular adaptation. *Mol. Biol. Evol.* 15, 355–369

32 Chang, B.S. and Donoghue, M.J. (2000) Recreating ancestral proteins. *Trends Ecol. Evol.* 15, 109–114

33 Fitch, W.M. *et al.* (1997) Long term trends in the evolution of H(3) HA1 human influenza type A. *Proc. Natl. Acad. Sci. U. S. A.* 94, 7712–7718

34 Bush, R.M. *et al.* (1999) Positive selection on the H3 hemagglutinin gene of human influenza virus A. *Mol. Biol. Evol.* 16, 1457–1465

35 Suzuki, Y. and Gojobori, T. (1999) A method for detecting positive selection at single amino acid sites. *Mol. Biol. Evol.* 16, 1315–1328

36 Nielsen, R. and Yang, Z. (1998) Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. *Genetics* 148, 929–936

37 Yang, Z. *et al.* (2000) Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics* 155, 431–449

38 Zanotto, P.M. *et al.* (1999) Genealogical evidence for positive selection in the *nef* gene of HIV-1. *Genetics* 153, 1077–1089

39 Yang, Z. *et al.* (2000) Maximum likelihood analysis of molecular adaptation in abalone sperm lysin reveals variable selective pressures among lineages and sites. *Mol. Biol. Evol.* 17, 1446–1455

40 Yamaguchi-Kabata, Y. and Gojobori, T. (2000) Reevaluation of amino acid variability of the human immunodeficiency virus type 1 gp120 envelope glycoprotein and prediction of new discontinuous epitopes. *J. Virol.* 74, 4335–4350

41 Bishop, J.G. *et al.* (2000) Rapid evolution in plant chitinases: molecular targets of selection in plant–pathogen coevolution. *Proc. Natl. Acad. Sci. U. S. A.* 97, 5322–5327

42 Haydon, D.T. *et al.* Evidence for positive selection in foot-and-mouth-disease virus capsid genes from field isolates. *Genetics* (in press)

43 Yang, Z. *et al.* (1998) Models of amino acid substitution and applications to mitochondrial protein evolution. *Mol. Biol. Evol.* 15, 1600–1611

44 Zhang, J. (2000) Rates of conservative and radical nonsynonymous nucleotide substitutions in mammalian nuclear genes. *J. Mol. Evol.* 50, 56–68

45 Riley, M.A. (1993) Positive selection for colicin diversity in bacteria. *Mol. Biol. Evol.* 10, 1048–1059

46 Hughes, A.L. and Yeager, M. (1997) Coordinated amino acid changes in the evolution of mammalian defensins. *J. Mol. Evol.* 44, 675–682

47 Qi, C.F. *et al.* (1998) Molecular phylogeney of *Fv1. Mamm. Genome* 9, 1049–1055

48 Tanaka, T. and Nei, M. (1989) Positive darwinian selection observed at the variable-region genes of immunoglobulins. *Mol. Biol. Evol.* 6, 447–459

49 Hughes, A.L. and Nei, M. (1988) Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. *Nature* 335, 167–170

50 Stotz, H.U. *et al.* (2000) Identification of target amino acids that affect interactions of fungal polygalacturonases and their plant inhibitors. *Mol. Physiol. Plant Path.* 56, 117–130

51 Kitano, T. *et al.* (1998) Conserved evolution of the *Rh50* gene compared to its homologous Rh blood group gene. *Biochem. Biophys. Res. Commun.* 249, 78–85

52 Zhang, J. *et al.* (1998) Positive darwinian selection after gene duplication in primate ribonuclease genes. *Proc. Natl. Acad. Sci. U. S. A.* 95, 3708–3713

53 Ford, M.J. *et al.* (1999) Natural selection promotes divergence of transferrin among salmonid species. *Mol. Ecol.* 8, 1055–1061

54 Hughes, A.L. (1995) The evolution of the type I interferon family in mammals. *J. Mol. Evol.* 41, 539–548

55 Goodwin, R.L. *et al.* (1996) Patterns of divergence during evolution of a_1-proteinase inhibitors in mammals. *Mol. Biol. Evol.* 13, 346–358

56 Hughes, M.K. and Hughes, A.L. (1995) Natural selection on *Plasmodium* surface proteins. *Mol. Biochem. Parasitol.* 71, 99–113

57 Wu, J.C. *et al.* (1999) Recombination of hepatitis D virus RNA sequences and its implications. *Mol. Biol. Evol.* 16, 1622–1632

58 Hughes, A.L. (1992) Positive selection and interallelic recombination at the merozoite surface antigen-1 (*MSA-1*) locus of *Plasmodium falciparum*. *Mol. Biol. Evol.* 9, 381–393

59 Smith, N.H. *et al.* (1995) Sequence evolution of the *porB* gene of *Neisseria gonorrhoeae* and *Neisseria meningitidis*: evidence for positive darwinian selection. *Mol. Biol. Evol.* 12, 363–370

60 Baric, R.S. *et al.* (1997) Episodic evolution mediates interspecific transfer of a murine coronavirus. *J. Virol.* 71, 1946–1955

61 Vacquier, V.D. *et al.* (1997) Positive darwinian selection on two homologous fertilization proteins: what is the selective pressure driving their divergence? *J. Mol. Evol.* 44, 15–22

62 Tsaur, S.C. and Wu, C-I. (1997) Positive selection and the molecular evolution of a gene of male reproduction, *Acp26Aa* of *Drosophila*. *Mol. Biol. Evol.* 14, 544–549

63 Karn, R.C. and Nachman, M.W. (1999) Reduced nucleotide variability at an androgen-binding protein locus (*Abpa*) in house mice: evidence for positive natural selection. *Mol. Biol. Evol.* 16, 1192–1197

64 Metz, E.C. and Palumbi, S.R. (1996) Positive selection and sequence arrangements generate extensive polymorphism in the gamete recognition protein bindin. *Mol. Biol. Evol.* 13, 397–406

65 Ting, C.T. *et al.* (1998) A rapidly evolving homeobox at the site of a hybrid sterility gene. *Science* 282, 1501–1504

66 Sutton, K.A. and Wilkinson, M.F. (1997) Rapid evolution of a homeodomain: evidence for positive selection. *J. Mol. Evol.* 45, 579–588

67 Rooney, A.P. and Zhang, J. (1999) Rapid evolution of a primate sperm protein: relaxation of functional constraint or positive darwinian selection? *Mol. Biol. Evol.* 16, 706–710

68 Ishimizu, T. *et al.* (1998) Identification of regions in which positive selection may operate in S-RNase of Rosaceae: implications for S-allele-specific recognition sites in S-RNase. *FEBS Lett.* 440, 337–342

69 Pamilo, P. and O'Neill, R.W. (1997) Evolution of *Sry* genes. *Mol. Biol. Evol.* 14, 49–55

70 Ward, T.J. *et al.* (1997) Nucleotide sequence evolution at the k-casein locus: evidence for positive selection within the family Bovidae. *Genetics* 147, 1863–1872

71 Duda, T.F., Jr and Palumbi, S.R. (1999) Molecular genetics of ecological diversification: duplication and rapid evolution of toxin genes of the venomous gastropod *Conus*. *Proc. Natl. Acad. Sci. U. S. A.* 96, 6820–6823

72 Nakashima, K. *et al.* (1995) Accelerated evolution in the protein-coding regions is universal in crotalinae snake venom gland phospholipase A_2 isozyme genes. *Proc. Natl. Acad. Sci. U. S. A.* 92, 5605–5609

73 Schmidt, T.R. *et al.* (1999) Molecular evolution of the COX7A gene family in primates. *Mol. Biol. Evol.* 16, 619–626
74 Wu, W. *et al.* (1997) Molecular evolution of cytochrome *c* oxidase subunit IV: evidence for positive selection in simian primates. *J. Mol. Evol.* 44, 477–491
75 Shields, D.C. *et al.* (1996) Evolution of hemopoietic ligands and their receptors: influence of positive selection on correlated replacements throughout ligand and receptor proteins. *J. Immunol.* 156, 1062–1070
76 Wallis, M. (1996) The molecular evolution of vertebrate growth hormones: a pattern of near-stasis interrupted by sustained bursts of rapid changes. *J. Mol. Evol.* 43, 93–100
77 Bargelloni, L. *et al.* (1998) Antarctic fish hemoglobins: evidence for adaptive evolution at subzero temperatures. *Proc. Natl. Acad. Sci. U. S. A.* 95, 8670–8675
78 Long, M. and Langley, C.H. (1993) Natural selection and the origin of *jingwei*, a chimeric processed functional gene in *Drosophila*. *Science* 260, 91–95
79 Muse, S.V. and Gaut, B.S. (1994) A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. *Mol. Biol. Evol.* 11, 715–724

Nice snake, shame about the legs

Michael Coates and Marcello Ruta

Snakes are one of the most extraordinary groups of terrestrial vertebrates, with numerous specializations distinguishing them from other squamates (lizards and their allies). Their musculoskeletal system allows creeping, burrowing, swimming and even gliding, and their predatory habits are aided by chemo- and thermoreceptors, an extraordinary degree of cranial kinesis and, sometimes, powerful venoms. Recent discoveries of indisputable early fossil snakes with posterior legs are generating intense debate about the evolutionary origin of these reptiles. New cladistic analyses dispute the precise significance and phylogenetic placement of these fossils. These conflicting hypotheses imply radically different scenarios of snake origins and relationships with wide biological implications.

Michael Coates and Marcello Ruta are at the Dept of Biology, Darwin Building, University College London, Gower Street, London, UK WC1E 6BT (m.coates@ucl.ac.uk; m.ruta@ucl.ac.uk).

The evolutionary origin of snakes (or Serpentes) has been discussed for over 130 years and their phylogenetic position within squamates is still debated. Around 2700 snake species are alive today and these are divided into three main groups[1–3] (Box 1): tiny fossorial (burrowing) scolecophidians (blindsnakes); anilioids (pipesnakes), which are mostly semi-fossorial; and macrostomatans, which include more familiar taxa, such as boas, pythons, vipers and cobras. In addition to the more obvious diagnostic characters of body elongation, limblessness and jaws that can engulf surprisingly large prey, other key features of snakes include absence of eyelids and external ears, and the presence of deeply forked tongues (linked to their highly attuned and sophisticated chemosensory systems[3]).

Ancestral diggers or swimmers?

Hypotheses concerning snake interrelationships fall into two main groups. For some researchers, snakes descend from terrestrial squamates that developed fossorial (burrowing) habits. Two groups of lizards exhibiting such habitats, amphisbaenians and dibamids, have often been regarded as snakes' closest living relatives[4]. Amphisbaenians, in particular, resemble scaly, loose-skinned earthworms, whose shovel-shaped or wedge-like heads function as soil-shunting devices. Specializations shared by snakes (Fig. 1a), amphisbaenians (Fig. 1b) and dibamids include loss, reduction and consolidation of skull bones; braincase enclosure; dorsal displacement of jaw-closing muscles; loss or reduction of limbs and girdles; and increased uniformity along the vertebral column. Furthermore, differences between the eyes of lizards and snakes are consistent with a model in which structures that were barely useful in a burrower underwent progressive reduction. Thus, whereas lizards, like humans, distort eye lens curvature to focus on objects, snakes lack ciliary muscles and are compelled to move the entire lens back and forth relative to the retina. Moreover, unlike lizards, snakes lack both a fovea and coloured oil droplets in retinal cells[1].

Alternative hypotheses[5] postulate that snakes are related to mosasauroids (Fig. 1c): spectacular marine reptiles from the upper half of the Cretaceous period, some 65–100 Mya[6]. Mosasauroids and snakes share reduced ossification of the pelvis and hindlimbs as well as specialized features of the jaw suspension and intramandibular joint kinetics (presence of a hinge allowing a degree of lateral movement within the lower jaw; Fig. 1a,c,d; Fig. 2, red circle). Phylogenetically, mosasauroids would be the nearest monophyletic sister group of snakes, with varanoid lizards (monitors) as the immediate sister group to this pair. Given this theory of relationships, the latest common ancestor of mosasaurs and snakes has been argued to have been a limbed, aquatic or semi-aquatic squamate[5,7–11]. Note that the implied ecological shift from an aquatic to a terrestrial environment in snake ancestry suggests that mosasaurs' (implied) aquatic habits were also primitive for Serpentes. Subsequently, snakes reduced and lost their limbs, although rudiments of the posterior pair remain in some forms, such as pythons.

Fossils: perfect missing links...

Renewed interest in the origin of snakes has been triggered by the recognition and discovery of three remarkable fossil forms with hind legs. Each of these ancient snakes is around 97 My old and originates from lowermost Upper Cretaceous sediments in the Middle East.

Pachyrhachis problematicus, from Israel (Fig. 1d–f), rapidly assumed a central position in debates about snake phylogeny[6,7,12]. It has miniature hindlimbs articulated with a rudimentary pelvic girdle (Fig. 1e,f), but sadly, its feet are missing. Currently described from only two specimens, it

 0169-5347/00/$ – see front matter PII: S0169-5347(00)01999-6