Inference of selection from multiple species alignments

Ziheng Yang

The selective pressure on a protein-coding gene can be measured by comparing silent (synonymous) and replacement (nonsynonymous) substitution rates. Higher replacement than silent rates provide unequivocal evidence for adaptive evolution driven by Darwinian selection. Previous employment of this criterion involved pairwise sequence comparison, averaging rates over time and sequences, resulting in virtually no power. Recent methods apply the criterion to particular lineages on a phylogeny or to individual sites in the gene and are much more powerful. Their application has led to detection of adaptive Darwinian selection in a number of genes and organisms.

Address
Galton Laboratory, Department of Biology, University College London,
Darwin Building, Gower Street, London WC1E 6BT, UK;
e-mail: z.yang@ucl.ac.uk

Current Opinion in Genetics & Development 2002, **12**:688–694

0959-437X/02/$ – see front matter

Published online 1 October 2002

Abbreviations

ARS	antigen recognition site
N	number of non-synonymous sites
S	number of synonymous sites
TPI	triosephosphate isomerase

Introduction

The relative importance of mutation and selection to molecular evolution has been a matter of debate for over thirty years. The neutral theory [1,2] claims that most observed molecular variation — both polymorphism within species and divergence between species — is not due to natural selection driving the fixation of advantageous mutations but to random fixation of selectively neutral mutations which confer no fitness advantage. Several tests of the neutral theory have been developed in population genetics [3,4,5••] and have been applied successfully to identify the footprints of positive selection from genome-wide analysis of within-species polymorphism [6,7]. Interpretations of such neutrality tests, however, are seldom unequivocal and often depend on assumptions about population demography and details of the selection model [5••,8].

A more stringent and robust criterion for detecting adaptive evolution in a protein-coding gene is an accelerated nonsynonymous (d_N, amino acid replacing) rate relative to the synonymous (d_S, silent) rate of substitution, with the rate ratio $\omega = d_N/d_S > 1$. As silent mutations do not change the amino acid whereas replacement mutations do, the difference in their fixation rates provides a measure of selective pressure on the protein. Because silent and replacement sites are interspersed in the same segment of the DNA, the effects of factors such as population size changes will be shared, providing a natural control [8,9]. Traditionally, synonymous and nonsynonymous substitution rates are defined in the context of comparing two DNA sequences, with d_N to be the number of nonsynonymous substitutions per nonsynonymous site and d_S to be the number of synonymous substitutions per synonymous site [10]. Thus, 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 will reduce its fixation rate so that $\omega < 1$. When the amino acid change offers a selective advantage, it will be fixed at a higher rate than a synonymous mutation, with $\omega > 1$. Purifying selection is rather easy to detect, and so the focus of my review is adaptive Darwinian selection indicated by $\omega > 1$. (See [11•] for a more mathematical treatment.)

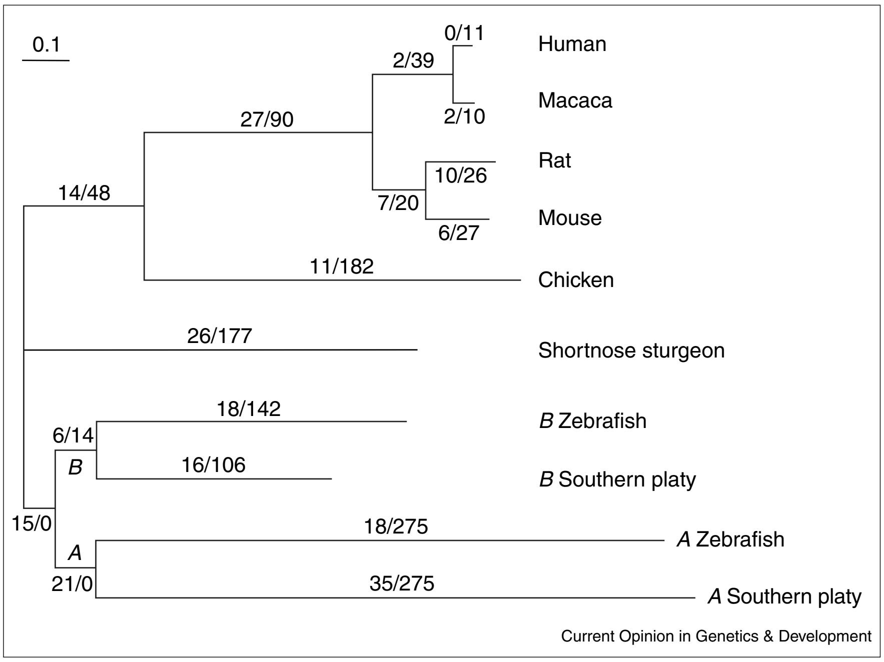
Early studies took an approach of pairwise sequence comparison and tested whether $d_N - d_S$ is significantly greater than 0. This approach in effect averages substitution rates over all amino acid sites in the sequence and over the time interval separating the two sequences. As most amino acid sites are expected to be highly conserved and adaptive evolution most likely affects only a few sites at a few time points [12], this approach has virtually no power to detect positive selection (e.g. [3]). For example, a large-scale database search identified only 17 proteins out of 3,595 as likely to be under positive selection, at a proportion of only 0.47% [13]. Indeed a very interesting use of such comparisons is to predict protein-coding regions in genomic DNA, making use of the fact that in virtually every gene, d_N is significantly smaller than d_S [14]. Recent methodological developments have largely remedied this problem of pairwise comparison.

Adaptive molecular evolution is at least partly responsible for evolutionary innovations and species divergences and is thus fundamentally important for understanding genome-scale evolution. Besides gaining a better understanding of forces and mechanisms of molecular evolution, identifying proteins or protein domains undergoing adaptive changes is also important for understanding gene function. Although it is well known that evolutionarily conserved regions of a protein are functionally important [15], it is not true that variable regions are all junk DNA. If the variability can be shown to be driven by positive Darwinian selection, functional importance is established.

Lineage-specific episodes of Darwinian selection

In an analysis of lysozyme evolution in primates, Messier and Stewart [16] inferred DNA sequences in extinct ancestral species and used them to calculate d_N and d_S for each branch in the phylogeny. Positive selection along a branch is identified by testing whether $d_N > d_S$, using a normal approximation to the statistic $d_N - d_S$. By focusing

Figure 1



Current Opinion in Genetics & Development

The phylogeny of the TPI genes. Branch *A* represents gene duplication leading to the new A isozyme. The unrooted tree is used in the analysis, although the root is most likely to be along the branch ancestral to chicken and mammals [22]. The branch lengths are measured by the expected number of nucleotide substitutions per codon, estimated under the free-ratio model which estimates one ω for each branch. The numbers along each branch are the likelihood estimates of nonsynonymous and synonymous changes (n^*/s^*) under the same model. Estimates under other models are listed in Table 1 for branch *A*.

on a single branch, the approach avoids averaging over long time periods, and it identified two lineages under positive selection.

Zhang *et al.* [17,18] were concerned about the reliability of the normal approximation and suggested the use of Fisher's exact test applied to counts of differences. (An example is presented later in this section.) This approach may suffer from several problems. First, reconstructed ancestral sequences are not real data and involve systematic biases and random errors [19]. Second, the methods used to estimate substitution rates along each branch are typically simplistic and do not account for features of sequence evolution such as different transition and transversion rates or codon usage bias. A common mistake made in using this approach is to apply the test to all branches in the phylogeny, in which case multiple tests are performed on the same data so that the significance values are incorrect.

The first two problems can be circumvented by taking a likelihood approach under a codon-substitution model [20,21]. In the likelihood calculation, the ω ratio can be assumed to be different among branches of the tree. Thus the hypothesis that a lineage of interest is under the same selective pressure as other lineages can be tested by comparing two models, one assuming the same ω ratio for all lineages and the other assuming a different ω for the lineages of interest [21]. Similarly, the likelihood ratio test can be used to test whether the ω ratio for the lineages of interest is significantly >1.

Here I use the data of Merritt and Quattro [22] from the triosephosphate isomerase (TPI) gene family in fishes. TPI is a dimeric glycolytic enzyme that catalyzes the interconversion of dihydroxyacetone phosphate and glyceraldehydes-3-phosphate and is central to carbohydrate metabolism. In all jawed vertebrates, a single neutrally-charged TPI protein is expressed. However, in teleost fish, two TPI proteins are found: the neutrally charged and generally expressed B form and the negatively charged derived A form that is found in neurons (eyes and brain). Merritt and Quattro [22] sequenced the TPI-coding cDNAs from two teleost fish species, southern platy and zebrafish, as well as the single gene (B form) in the more primitive fish shortnose sturgeon (see Figure 1). The authors performed an analysis of adaptive evolution after gene duplication along branch *A* in Figure 1 using the method of Messier and Stewart [16] and Zhang *et al.* [17]. The empirical Bayes approach of Yang *et al.* [19] was used to reconstruct amino acid sequences for ancestral nodes, with branch lengths estimated using least squares [17]. The coding DNA sequences for ancestral nodes were then inferred according to the reconstructed amino acid sequences. The reconstruction suggested $n = 23$ nonsynonymous changes and $s = 1$ synonymous change along branch *A*, whereas the numbers of nonsynonymous and synonymous sites in the sequence were calculated to be $N = 536$ and $S = 196$. Use of Fisher's exact test [17] suggested a significant excess of nonsynonymous changes along branch *A*, with $P = 0.005$.

To illustrate the bias in ancestral sequence reconstruction and to examine the sensitivity of the analysis to model assumptions, I analyze the same data using maximum likelihood under models of codon substitution [21]. The

Table 1

Maximum likelihood estimates of d_N and d_S for branch *A* in the TPI phylogeny.

	Maximum likelihood estimates						Reconstruction
Model	p	ℓ	κ	$\omega_A = d_N/d_S$	n^*/s^*	N/S	n/s
(A) 1-ratio, Fequal, κ=1	18	−4,129.09	1	0.070 = 0.026/0.373	14.4/70.1	550.0/188.0	21.5/35.5
(B) 1-ratio, Fequal	19	−4,122.00	1.4	0.076 = 0.026/0.346	14.3/68.0	541.8/196.2	21.5/37.5
(C) 1-ratio, F3x4	28	−4,097.58	1.3	0.054 = 0.029/0.536	16.0/96.8	557.5/180.5	21.5/41.5
(D) 3-ratios, Fequal	21	−4,116.79	1.4	0.425 = 0.038/0.090	20.8/17.7	542.0/196.0	21.5/9.5
(E) 3-ratios, F3x4	30	−4,094.21	1.5	0.195 = 0.037/0.153	20.4/27.6	557.8/180.2	21.5/17.5
(F) Free-ratios, Fequal	35	−4,100.54	1.4	∞ = 0.039/0.000	20.9/0.0	541.9/196.1	22.0/0.0
(G) Free-ratios, F3x4	44	−4,076.95	1.5	∞ = 0.038/0.000	21.1/0.0	557.7/180.3	23.0/0.0
(H) Free-ratios, F3x4, ω_A = 1	43	−4,077.00	1.5	1 = 0.038/0.038	20.9/6.8	557.7/180.3	23.0/1.0

NB, p is the number of parameters in the model including 17 branch lengths (Figure 1) and κ is the transition/transversion rate ratio. The numbers of nonsynonymous and synonymous changes along branch *A* are calculated as $n^* = d_N \times N$ and $s^* = d_S \times S$, where N and S are the numbers of nonsynonymous and synonymous sites, respectively. n and s are counts of differences from the reconstructed ancestral sequences, which are inferred using the Empirical Bayes method [19] under the codon model. The Fequal model ignores codon usage bias and assumes equal frequency for each sense codon (1/61), whereas F3x4 calculates expected codon frequencies using the nucleotide frequencies at the three codon positions.

results are shown in Table 1. The one-ratio models (A-C) assume the same ω ratio for all branches and are used to demonstrate the effects of transition/transversion bias and codon usage bias on calculation of the numbers of synonymous (S) and nonsynonymous (N) sites. Ignoring the transition/transversion rate bias is well known to cause an underestimate of synonymous sites S [23]. Ignoring base composition bias has the opposite effect, and leads to overestimation of S [24]. The base compositions at the third codon position are 21% for T, 31% for C, 14% for A, and 33% for G. The high GC content means that most changes at the third position are transversions between G and C, which are more likely to be nonsynonymous than are random changes between the four nucleotides. When both biases are accounted for, the expected numbers of sites are $N = 557.8$ and $S = 180.2$.

The one-ratio models also highlight the differences between the likelihood estimates of changes (n^* and s^*) and the counts from the reconstructed ancestral sequences (n and s), even if both are from the same analysis. For example, when no codon usage bias and no transition/transversion bias are assumed (model A, Table 1), the likelihood estimates suggest about $n^* = 14$ nonsynonymous substitutions and $s^* = 70$ synonymous substitutions along branch A in the tree of Figure 1, whereas the corresponding numbers from ancestral reconstruction are $n = 21$ and $s = 35$. Similarly, large differences are observed for other branches in the tree (Table 1). The reconstruction approach does not correct for multiple hits within the branch, and might be expected to underestimate the number of substitutions. Although the method indeed underestimates synonymous changes, it surprisingly overestimates nonsynonymous changes. Thus, counts of differences (n and s) from reconstructed sequences are unreliable pseudo-data whereas counts of sites (N and S) are affected by codon usage bias; as a result, Fisher's exact test based on those counts is not exact.

To test for positive directional selection along branch A, I consider two models. The 'three-ratios' model assumes two independent ω ratios for branches A and B (Figure 1), whereas all other branches have the same ratio ω_0. The 'free-ratios' model assumes an independent ω for each branch on the phylogeny. Estimates of ω_A under the three-ratios models are several times higher than ω_0 for other branches. Nevertheless, the estimates are less than one, and thus the evidence for positive selection is not clear-cut (Table 1). Under the free-ratios model, ω_A is estimated to be infinity, with 21 nonsynonymous changes and 0 synonymous changes along the branch. However, fitting the free-ratios model with $\omega_A = 1$ fixed (last row, Table 1) reduces the log likelihood by only 0.5. Thus the null hypothesis $\omega_A = 1$ cannot be rejected, with $P = 0.75$.

The difference in the ω_A estimate between the 3-ratios and free-ratios models is surprising as both models allow ω_A to vary freely. In previous analyses, those models were noted to produce similar estimates for the common parameters (e.g. [21,25]). However, the TPI gene sequences are quite divergent and at some codon sites, all three positions are different either between the zebrafish and southern platy at the A locus or between the A and B loci. Ancestral reconstruction at those sites is unreliable, especially concerning distributions of synonymous changes. The sensitivity of maximum likelihood estimates to model assumptions indicates a similar lack of information in the data. Thus, the statistical support from the sequence data for positive selection is not unequivocal, and the accelerated nonsynonymous rate along branch A is compatible with both positive selection and relaxation of functional constraints (purifying selection) (Table 1). However, both the A and B forms of TPI are clearly functional, and their biochemical differences [22] suggest that positive selection is a more likely explanation. Sequencing more teleost fish species to break the long branches in the tree will likely settle the issue.

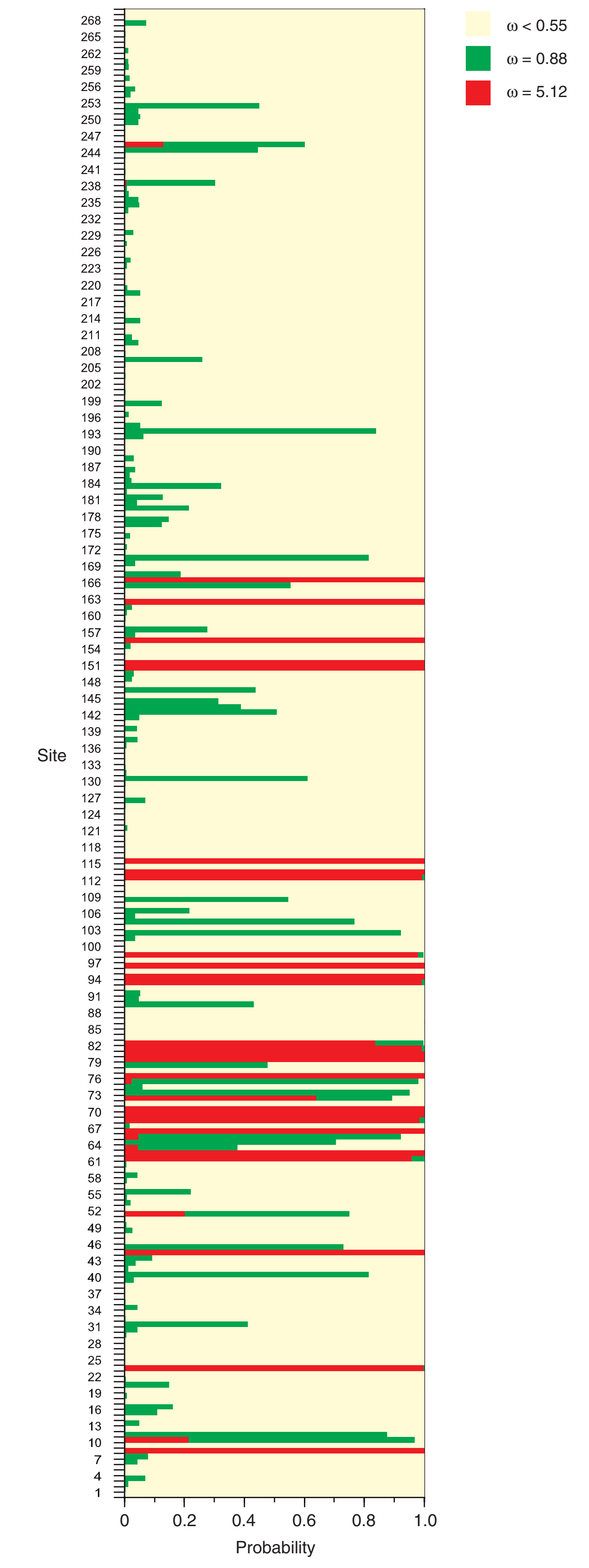
Amino acid sites under Darwinian selection

The lineage-based analyses discussed above assume that all amino acid sites are under the same selective pressure

Figure 2 legend

Posterior probabilities of site classes for sites along the MHC class I gene. A dataset of 192 alleles from the human class I MHC alleles was analysed under the random-sites model M8 (beta&ω). Maximum likelihood parameter estimates suggest 90.0% of conserved sites with ω ratios from the β distribution $B(p = 0.168, q = 0.710)$ and 10.0% of positive selection sites with $\omega = 5.122$. Ten equal-probability categories are used to approximate the β distribution [31], with ω ratios of 0.000, 0.000, 0.000, 0.003, 0.015, 0.048, 0.128, 0.286, 0.548, 0.881, and 5.122. The first nine categories are collapsed into one category represented by $\omega < 0.55$. Site numbering is according to the structure file 1AKJ in Protein Data Bank (chain A). From [27•].

Figure 2



Current Opinion in Genetics & Development

and effectively average the ω ratio across all sites. Positive selection is detected for a lineage only if that average is greater than one. This is a very 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.

If prior information is available about which amino acid residues are potentially under positive selection, attention can be focused on them. The best-known example is Hughes and Nei's [26] analysis of the MHC alleles, where the protein 3-D structure was used to identify amino acid residues at the antigen recognition site (ARS). Positive selection was detected for those sites but not for the whole gene. Those results were confirmed in a recent likelihood analysis, using so-called fixed-sites models that assign and estimate separate ω ratios for ARS and non-ARS codons [27•].

When such information is unavailable, several approaches can still be used. One is the sliding window analysis (e.g. [13]). The d_N and d_S rates are calculated for a sliding window and regions with higher d_N than d_S can be identified. Even this method often lacks power, as positively selected sites may be interspersed among highly conserved sites in the sequence [28]. It would be more sensible if we could slide the window on the protein tertiary structure rather than along the primary sequence. Thus the focus has been on methods designed to detect selection at individual sites. Fitch *et al.* [29] 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 the sequence. Suzuki and Gojobori [28] took a similar approach. They estimated the numbers of synonymous and nonsynonymous sites and differences along the tree using reconstructed ancestral sequences at each site. They then tested whether the proportion of nonsynonymous substitutions differed from the neutral expectation ($\omega = 1$). When the data contain many similar sequences, these methods are useful tools for exploratory data analysis but the use of reconstructed ancestral sequences and the simplistic assumptions made in the counting of sites and differences make hypothesis testing unreliable [11•].

In the likelihood method, it is not advisable to estimate one ω ratio parameter for each site. Instead, a statistical distribution is used to describe the variation of ω among sites [30,31]. Since ω is a measure of selective pressure on the protein, these models account for variable selective pressures among sites. However, such random-sites models make no prior assumption about which sites might be under positive selection. To test for positive selection, that is, the presence of sites at which $\omega > 1$, we compare two statistical distributions. The null model uses a distribution that does not allow for sites with $\omega > 1$ whereas the more-general alternative model does. One such pair consists of models M7 (beta) and M8 (beta&ω) [31]. The beta distribution (M7) is rather flexible but is limited to the range $0 < \omega < 1$ and thus serves as the null model. Model M8 adds an additional site class that has an ω freely estimated from the data and can thus be >1. The null and alternative models can be compared by a likelihood-ratio test to see whether allowing for sites under positive selection provides a significant improvement to the fit of the model to data. Over a dozen such distributions were implemented [31].

When parameter estimates under the alternative model (such as M8) suggest the presence of sites with $\omega > 1$, Bayes theorem can be used to calculate the posterior probability that each site is from that site class, and sites having high probabilities are potential targets of positive selection. An example is shown in Figure 2, where posterior probabilities are plotted along the MHC sequence, highlighting sites that are potentially under diversifying selection [27•]. The model identified 25 sites to be under positive selection at the 95% probability level, out of which 22 are in the antigen recognition site while 3 are nearby. The model was able to distinguish which of the 57 amino acid residues at the antigen recognition site were truly under diversifying selection and which of them are actually highly conserved. These studies appear very useful in generating hypotheses for laboratory investigation because they could identify crucial amino acids whose changes have offered a selective advantage in Nature's grand evolutionary experiment.

Note that the likelihood ratio test answers the simple question of whether the sequence has sites under positive selection with $\omega > 1$, while Bayes calculation of posterior probabilities is used to pinpoint such sites. The latter is clearly a much more difficult task than the former as there may not be enough information to allow inference at every site. Because the information for such inference comes from synonymous and nonsynonymous changes at individual sites, the most important factors affecting performance are the number of sequences and the level of sequence divergence. Those intuitive expectations have been confirmed by computer simulations [32,33•]. Indeed, the likelihood ratio test was found to perform well even in small datasets [32]. By contrast, Bayes identification of those sites under positive selection performed poorly when the data contained a few highly similar sequences although performance increased dramatically with the inclusion of more sequences and more sequence divergence. Furthermore, both the likelihood ratio test of positive selection and the Bayes identification of sites under selection were found to be highly inaccurate for within-species data due to lack of variability. In population data, recombination is also a concern, the effects of which are not yet well understood [33•].

While the likelihood method [30,31] enjoys a better statistical justification than methods based on ancestral reconstruction [28,29], it should be emphasized that both methods make use of similar information in the data and require multiple moderately divergent sequences for reliable inference. In large datasets with many sequences, the two methods may be expected to locate similar sites [34,35]. In smaller datasets, there may be insufficient changes at any single site to suggest selection. The likelihood ratio test will then have more power to detect the presence of positive selection as it combines evidence over many sites.

Application and performance in real data

The methods reviewed here appear to be conservative. They detect selection only if d_N is higher than d_S, and selection that does not cause excessive replacement substitutions may not be detected. 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, tests of positive selection at sites work only if the ω ratio averaged over all branches is >1. Adaptive evolution that occurs only during a short time interval and that affects only a few crucial amino acids may not be detected by any of these methods.

In this regard the success of these methods has been remarkable. The lineage-based analysis has been used to detect positive selection indicated by a burst of replacement substitutions along particular lineages of a phylogeny. Recent examples include directional selection in pheromone-binding proteins along moth lineages in which pheromone changes have occurred [36]; adaptive evolution in the breast cancer gene BRCA1 in the human and chimpanzee lineages [37]; adaptive evolution of hemoglobins in Antarctic teleost fish (Notothenioids) at subzero temperature [38]; and positive selection in transferrin, an iron-binding protein, in salmonids, probably due to competition for iron from pathogenic bacteria [39]. The lineage-based methods are particularly useful in analysis of gene-family evolution to demonstrate the role of positive selection driving functional divergence after gene duplication [18,25,40,41].

Nevertheless, averaging substitution rates over sites appears to be a more serious problem than averaging over lineages. As a result, models accounting for variable selective pressures among sites [30,31] appear even more successful, detecting positive selection even in a background of overwhelming purifying selection. A large number of genes involved in host-pathogen or host-parasite antagonism have thus been identified to contain sites under positive selection [31,42–47,48•,49]. In such systems, an evolutionary arms race appears to be going on, causing accelerated replacement substitution rates at amino acid residues involved in the recognition in either the host or the pathogen or parasite. An interesting example concerns the evolution of the *wsp* gene among Wolbachia bacteria, thought to be involved in host-parasite interactions. The bacterial gene was found to be under diversifying selection when the bacteria have an antagonistic relationship with its hosts (arthropods) while not when the bacteria have a mutualistic relationship with its hosts (nematode worms) [50]. Several studies have also identified positive selection involved in sperm-egg recognition [51] or other aspects of male or female reproduction [52–56], providing molecular evidence for the evolutionary battle of the sexes.

Conclusions

An accelerated replacement rate over silent rate is a very stringent criterion for detecting adaptive molecular evolution. Yet, methods focused on selection in a short period of time or at individual sites have enjoyed much success when applied to real data analysis. With the fast accumulation of DNA sequences, more exciting cases of adaptive molecular evolution are bound to be found. We are now at a position to search genome-wide for molecular Darwinism and to answer questions such as what proportion of proteins are evolving under the influence of positive Darwinian selection. Such studies not only advance our understanding of the mechanisms of molecular evolution but also provide insights into protein function.

Acknowledgements

I thank Thomas Merritt for comments. This work is supported by grants from the Biotechnology and Biological Sciences Research Council (UK) and the Human Frontier Science Programme (EU).

References and recommended reading

Papers of particular interest, published within the annual period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Kimura M: **Evolutionary rate at the molecular level**. *Nature* 1968, **217**:624-626.

2. King CE, Jukes TH: **Non-Darwinian evolution**. *Science* 1969, **164**:788-798.

3. Kreitman M, Akashi H: **Molecular evidence for natural selection**. *Annu Rev Ecol Syst* 1995, **26**:403-422.

4. Wayne ML, Simonsen KL: **Statistical tests of neutrality in the age of weak selection**. *Trends Ecol Evol* 1998, **13**:236-240.

5. •• Fay JC, Wu C-I: **The neutral theory in the genomic era**. *Curr Opin Genet Dev* 2001, **11**:642-646.

An exciting review of neutrality tests and their uses for analysing genomic data to detect the footprints of natural selection.

6. Fay JC, Wyckoff GJ, Wu CI: **Testing the neutral theory of molecular evolution with genomic data from Drosophila**. *Nature* 2002, **415**:1024-1026.

7. Smith NG, Eyre-Walker A: **Adaptive protein evolution in Drosophila**. *Nature* 2002, **415**:1022-1024.

8. Nielsen R: **Statistical tests of selective neutrality in the age of genomics**. *Heredity* 2001, **86**:641-647.

9. Akashi H: **Inferring the fitness effects of DNA mutations from polymorphism and divergence data: statistical power to detect directional selection under stationarity and free recombination**. *Genetics* 1999, **151**:221-238.

10. Miyata T, Yasunaga T: **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* 1980, **16**:23-36.

11. • Yang Z: **Adaptive molecular evolution.** In *Handbook of Statistical Genetics*. Edited by Balding D, Bishop M, Cannings C. New York: Wiley; 2001:327-350.

In this book chapter, I provide a more theoretical treatment of most of the methods discussed here. See also the PAML (for Phylogenetic Analysis by Maximum Likelihood) programme web page (http://abacus.gene.ucl.ac.uk/software/paml.html) for implementation details.

12. Gillespie JH: *The Causes Of Molecular Evolution*. Oxford: Oxford University Press; 1991.

13. Endo T, Ikeo K, Gojobori T: **Large-scale search for genes on which positive selection may operate.** *Mol Biol Evol* 1996, **13**:685-690.

14. Nekrutenko A, Makova KD, Li W-H: **The K^A/K^S ratio test for assessing the protein-coding potential of genomic regions: an empirical and simulation study.** *Genome Res* 2001, **12**:198-202.

15. Li W-H: *Molecular Evolution*. Massachusetts, USA: Sinauer Associates; 1997.

16. Messier W, Stewart C-B: **Episodic adaptive evolution of primate lysozymes**. *Nature* 1997, **385**:151-154.

17. Zhang J, Kumar S, Nei M: **Small-sample tests of episodic adaptive evolution: a case study of primate lysozymes**. *Mol Biol Evol* 1997, **14**:1335-1338.

18. Zhang J, Rosenberg HF, Nei M: **Positive Darwinian selection after gene duplication in primate ribonuclease genes**. *Proc Natl Acad Sci USA* 1998, **95**:3708-3713.

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

20. Yang Z, Nielsen R: **Synonymous and nonsynonymous rate variation in nuclear genes of mammals**. *J Mol Evol* 1998, **46**:409-418.

21. Yang Z: **Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution**. *Mol Biol Evol* 1998, **15**:568-573.

22. Merritt TJ, Quattro JM: **Evidence for a period of directional selection following gene duplication in a neurally expressed locus of triosephosphate isomerase**. *Genetics* 2001, **159**:689-697.

23. Li W-H, Wu C-I, Luo C-C: **A new method for estimating synonymous and nonsynonymous rates of nucleotide substitutions considering the relative likelihood of nucleotide and codon changes**. *Mol Biol Evol* 1985, **2**:150-174.

24. Yang Z, Nielsen R: **Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models**. *Mol Biol Evol* 2000, **17**:32-43.

25. Bielawski JP, Yang Z: **Positive and negative selection in the DAZ gene family**. *Mol Biol Evol* 2001, **18**:523-529.

26. Hughes AL, Nei M: **Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection**. *Nature* 1988, **335**:167-170.

27. • Yang Z, Swanson WJ: **Codon-substitution models to detect adaptive evolution that account for heterogeneous selective pressures among site classes**. *Mol Biol Evol* 2002, **19**:49-57.

This paper uses a dataset of 192 class I MHC alleles from humans to compare the fixed-sites models, which use structural information to partition the ARS and non-ARS codons, and the random-sites models. The analysis demonstrated the surprising power of the random-sites fishing-expedition models, which were able to distinguish codons at the ARS truly under positive selection from those that are actually highly conserved.

28. Suzuki Y, Gojobori T: **A method for detecting positive selection at single amino acid sites**. *Mol Biol Evol* 1999, **16**:1315-1328.

29. Fitch WM, Bush RM, Bender CA, Cox NJ: **Long term trends in the evolution of H(3) HA1 human influenza type A**. *Proc Natl Acad Sci USA* 1997, **94**:7712-7718.

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

31. Yang Z, Nielsen R, Goldman N: **Pedersen A-MK: Codon-substitution models for heterogeneous selection pressure at amino acid sites**. *Genetics* 2000, **155**:431-449.

32. Anisimova M, Bielawski JP, Yang Z: **The accuracy and power of likelihood ratio tests to detect positive selection at amino acid sites**. *Mol Biol Evol* 2001, **18**:1585-1592.

33. Anisimova M, Bielawski JP, Yang Z: **Accuracy and power of Bayes**
• **prediction of amino acid sites under positive selection**. *Mol Biol Evol* 2002, **19**:950-958.

Computer simulation was conducted to investigate the effects of factors such as the size of the tree and sequence divergence on the performance of the likelihood method for detecting positively selected sites. The paper provides practical guidelines about the feasibility of such inference.

34. Yang Z: **Maximum likelihood analysis of adaptive evolution in HIV-1 gp120 *env* gene**. *Pacific Symp Biocomp* 2001:226-237.

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

36. Willett CS: **Evidence for directional selection acting on pheromone-binding proteins in the genus Choristoneura**. *Mol Biol Evol* 2000, **17**:553-562.

37. Huttley GA, Easteal S, Southey MC, Tesoriero A, Giles GG, McCredie MRE, Hopper JL, Venter DJ: **Adaptive evolution of the tumour suppressor BRCA1 in humans and chimpanzees**. *Nat Genet* 2000, **25**:410-413.

38. Bargelloni L, Marcato S, Patarnello T: **Antarctic fish hemoglobins: evidence for adaptive evolution at subzero temperatures**. *Proc Natl Acad Sci USA* 1998, **95**:8670-8675.

39. Ford MJ: **Molecular evolution of transferrin: evidence for positive selection in salmonids**. *Mol Biol Evol* 2001, **18**:639-647.

40. Johnson ME, Viggiano L, Bailey JA, Abdul-Rauf M, Goodwin G, Rocchi M, Eichler EE: **Positive selection of a gene family during the emergence of humans and African apes**. *Nature* 2001, **413**:514-519.

41. Maston GA, Ruvolo M: **Chorionic gonadotropin has a recent origin within primates and an evolutionary history of selection**. *Mol Biol Evol* 2002, **19**:320-335.

42. Baum J, Ward R, Conway D: **Natural selection on the erythrocyte surface**. *Mol Biol Evol* 2002, **19**:223-229.

43. Zanotto PM, Kallas EG, Souza RF, Holmes EC: **Genealogical evidence for positive selection in the *nef* gene of HIV-1**. *Genetics* 1999, **153**:1077-1089.

44. Bishop JG, Dean AM, Mitchell-Olds T: **Rapid evolution in plant chitinases: molecular targets of selection in plant-pathogen coevolution**. *Proc Natl Acad Sci USA* 2000, **97**:5322-5327.

45. Haydon DT, Bastos AD, Knowles NJ, Samuel AR: **Evidence for positive selection in foot-and-mouth-disease virus capsid genes from field isolates**. *Genetics* 2001, **157**:7-15.

46. Jenkins GM, Worobey M, Woelk CH, Holmes EC: **Non-quasispecies evidence for the evolution of RNA Viruses**. *Mol Biol Evol* 2001, **18**:987-994.

47. Fares MA, Moya A, Escarmis C, Baranowski E, Domingo E, Barrio E: **Evidence for positive selection in the capsid protein-coding region of the foot-and-mouth disease virus (FMDV) subjected to experimental passage regimens**. *Mol Biol Evol* 2001, **18**:10-21.

48. Stahl EA, Bishop JG: **Plant-pathogen arms races at the molecular**
• **level**. *Curr Opin Plant Biol* 2000, **3**:299-304.

The review provides a summary of evidence for and mechanisms of molecular adaptation in both plants and pathogens involved in an evolutionary arms race.

49. Yang Z, Bielawski JP: **Statistical methods for detecting molecular adaptation**. *Trends Ecol Evol* 2000, **15**:496-503.

50. Jiggins FM, Hurst GDD, Yang Z: **Host-symbiont conflicts: positive selection on the outer membrane protein of parasite but not mutualistic Rickettsiaceae**. *Mol Biol Evol* 2002, **19**:1341-1349.

51. Palumbi SR: **Genetic divergence, reproductive isolation and marine speciation**. *Annu Rev Ecol Syst* 1994, **25**:547-572.

52. Tsaur SC, Wu C-I: **Positive selection and the molecular evolution of a gene of male reproduction, *Acp26Aa* of *Drosophila***. *Mol Biol Evol* 1997, **14**:544-549.

53. Hellberg ME, Vacquier VD: **Positive selection and propeptide repeats promote rapid interspecific divergence of a gastropod sperm protein**. *Mol Biol Evol* 2000, **17**:458-466.

54. Wyckoff GJ, Wang W, Wu C-I: **Rapid evolution of male reproductive genes in the descent of man**. *Nature* 2000, **403**:304-309.

55. Swanson WJ, Yang Z, Wolfner MF, Aquadro CF: **Positive Darwinian selection in the evolution of mammalian female reproductive proteins**. *Proc Natl Acad Sci USA* 2001, **98**:2509-2514.

56. Swanson WJ, Clark AG, Waldrip-Dail HM, Wolfner MF, Aquadro CF: **Evolutionary EST analysis identifies rapidly evolving male reproductive proteins in *Drosophila***. *Proc Natl Acad Sci USA* 2001, **98**:7375-7379.