

## SHORT REVIEW

# The quest for natural selection in the age of comparative genomics

M Anisimova<sup>1</sup> and DA Liberles<sup>2</sup>

<sup>1</sup>Department of Biology, University College London, London, UK and <sup>2</sup>Department of Molecular Biology, University of Wyoming, Laramie, WY, USA

Continued genome sequencing has fueled progress in statistical methods for understanding the action of natural selection at the molecular level. This article reviews various statistical techniques (and their applicability) for detecting adaptation events and the functional divergence of proteins. As large-scale automated studies become more frequent, they provide a useful resource for generating biological null hypotheses for further experimental and statistical testing. Furthermore, they shed light on typical patterns of lineage-specific evolution of organisms, on the functional and

structural evolution of protein families and on the interplay between the two. More complex models are being developed to better reflect the underlying biological and chemical processes and to complement simpler statistical models. Linking molecular processes to their statistical signatures in genomes can be demanding, and the proper application of statistical models is discussed.

*Heredity* advance online publication, 12 September 2007; doi:10.1038/sj.hdy.6801052

**Keywords:** functional divergence; positive selection; comparative genomics; molecular evolution; protein structure; population genetics

## Introduction

Genomic change underlies the biodiversity found on Earth. Rapid genome sequencing coupled with the development of statistical methods for comparative genomics has enabled the examination of forces driving lineage-specific divergence at the molecular level. In an early study focusing on a comparison of human and chimpanzee, the low levels of sequence divergence suggested the hypothesis that evolutionary changes in gene regulation have driven phenotypic divergence between species (King and Wilson, 1975). More recently, the rewiring of a regulatory pathway from the last common ancestor of the sea urchin and starfish toward both extant species (Hinman *et al.*, 2003) and the evolution of pigmentation in *Drosophila* (Prud'homme *et al.*, 2006) gave further support to this view. Although changes in coding sequences can be compensatory (Haag, 2007), changes at the protein-coding level clearly also play a vital role in phenotypic divergence, as experimental evidence of lineage-specific functional change of proteins has been found in a growing number of cases (Benner *et al.*, 2007). Detecting genes targeted by selection in the genome became an efficient strategy for finding causes of species differences and identifying genomic regions of functional, and potentially medical, significance. Rather than focus on the controversy surrounding the genomic basis of phenotypic evolution, we will assume that evolution of protein-coding

sequences contributes to changes in the phenotype and fitness of the organism and proceed with a discussion of computational methods to detect selection-driven changes leading to changes in molecular phenotype (protein function).

Several views exist of how gene sequence maps to protein function. From a gradualist viewpoint, proteins accumulate substitutions and this gradual change corresponds to a steady evolution of protein function. This is largely consistent with a neutralist view of protein evolution, in which functional change is not being driven by positive selection and is due to a random accumulation of mutations, but this scenario is now commonly incorporated into a selectionist viewpoint. With bursts of lineage-specific positive selection, punctuated periods of rapid sequence change may occur on a few branches (Gould and Eldredge, 1993), whereas negative selection and functional stasis are seen on others (Messier and Stewart, 1997). Both gradualist and punctuated views of protein sequence evolution are consistent with protein structure dictating sites where substitution can occur and those where any change would radically diminish protein fitness, on the basis of both binding interactions and folding constraints (Bloom *et al.*, 2007; Lin *et al.*, 2007). Functionally important sites are expected to evolve slowly, whereas rapid changes are expected at sites that have little impact on the structure and function of the protein (neutral positions) or at sites where diversification is favored through increased fitness of the organism or necessitated as compensatory changes (Depristo *et al.*, 2005; Lin *et al.*, 2007). Such heterogeneity across sites is frequently modeled with a  $\gamma$  distribution (Yang, 1996; Stern and Pupko, 2006). Changes to protein function can be altered not just by changes in specific amino acids

Correspondence: Dr DA Liberles, Department of Molecular Biology, University of Wyoming, Department 3944, Laramie, WY 82071, USA.  
E-mail: liberles@uwyo.edu  
Received 30 April 2007; revised 19 June 2007; accepted 3 August 2007

(positive selection), but also by the substitution rates at some sites as a result of changes in functional and structural constraints at specific residues (Gaucher *et al.*, 2002). Debate is currently open on whether detection of such shifts of selective pressures at individual sites frequently occurs neutrally (Lopez *et al.*, 2002) or is linked to functional divergence (Abhiman and Sonnhammer, 2005b; Gu *et al.*, 2007). A further question is whether changes in selective pressure are directed by coevolutionary processes dictated by structure (Pollock *et al.*, 1999; Suzuki, 2004; Berglund *et al.*, 2005; Parisi and Echave, 2005) or are occurring neutrally and independently from protein structure (Galtier and Jean-Marie, 2004; Dutheil *et al.*, 2005). This depends on the amount of selection dictated by function compared with structural compensatory evolution without functional shifts (Depristo *et al.*, 2005).

Underlying questions of gradualism versus punctuated behavior includes what the fitness landscape looks like and how easily a new function can evolve. Golding and Dean (1998) pointed to cases in which adaptive functional change can occur with only a single point substitution. Consistent with a punctuated view, functional adaptation may be driven by single changes or by small numbers of changes. This would make such events difficult to detect statistically. Although a small number of substitutions can alter function, more changes might be expected to be seen as a result of several processes. These processes include coevolving residues that optimize a new function and selective sweeps bringing linked nonadaptive changes to fixation. From a gradualist viewpoint, the power to detect lineage-specific changes may be additionally weakened by a slower response to selection and a dilution of selected residues by those evolving at the neutral background.

Discussion about the way proteins evolved has been reflected in the development of computational methods incorporating new angles of this biological complexity. We dedicate the remaining space to reviewing such methods and their applicability at a comparative genomic level.

## Detecting selection at the population genetic level

Many methods have been proposed for population data. Tajima's  $D$ -test (for DNA data) compares the estimate of the population-scaled mutation rate based on the number of pairwise differences with that based on the number of segregating sites in a sample (Tajima, 1989). Under neutrality, Tajima's  $D \approx 0$ , and significant deviations may indicate a selective sweep ( $D < 0$ ) or balancing selection ( $D > 0$ ). Several other neutrality tests, based on slightly different summary statistics, use a similar idea (Fu and Li, 1993; Fay and Wu, 2000). The Hudson–Kreitman–Aguade test for DNA data evaluates the neutral hypothesis by comparing variability within and between species for two or more loci (Hudson *et al.*, 1987). Under neutrality, levels of polymorphism (variability within species) and divergence (variability between species) should be proportional to the mutation rate, resulting in a constant polymorphism-to-divergence ratio. Tests of selective neutrality based solely on simple summary statistics seem to be powerful enough to reject the strictly neutral model but are sensitive to

demographic assumptions (constant population size, absence of population structure and migration), making it difficult to obtain unambiguous evidence for selection (Wayne and Simonsen, 1998; Nielsen, 2001). The McDonald–Kreitman test for protein-coding data has been more successful at detecting selection (McDonald and Kreitman, 1991). Exploiting the underlying idea of the Hudson–Kreitman–Aguade test, the McDonald–Kreitman test compares the ratio of nonsynonymous (amino acid altering) to synonymous (silent) substitutions within and between species, which should be the same in the absence of selection. This test is more robust to demographic assumptions, as the effect of the demographic model should be the same for both nonsynonymous and synonymous sites (Nielsen, 2001).

However, neutrality tests do not distinguish between different forms of natural selection and so cannot provide explicit evidence for adaptive evolution (Yang and Bielawski, 2000). Various modifications of the McDonald–Kreitman test (Templeton, 1996; Akashi, 1995, 1999b) proved more informative about the nature of selective forces. In particular, Akashi suggested a way to differentiate between the types of selection by examining the frequency distribution of observed silent and replacement changes compared with the neutral expectation. However, the power of the test is low when selection is weak or the fraction of adaptive mutations is small. Moreover, deviations from neutrality may be equally attributed to a population expansion or bottleneck (Eyre-Walker, 2002; Smith and Eyre-Walker, 2002). Whereas the population demographic process is expected to affect all genomic loci, selection should be nonuniform. Several studies (Eyre-Walker, 2002; Fay *et al.*, 2001, 2002; Smith and Eyre-Walker, 2002) took a genomic approach and confirmed that polymorphism to divergence ratios differed significantly only for a few genes, although the high amounts of inferred adaptation exceeded expectations.

Unlike neutrality tests that do not explicitly model selection, the Poisson random-field framework (Sawyer and Hartl, 1992; Hartl *et al.*, 1994; Akashi, 1999a) enables estimation of mutation and selection parameters in various population genetics scenarios. The rationale behind the approach is that natural selection alters the site-frequency spectrum, making it possible to estimate the strength of selection that has contributed to the observed deviation from neutrality. On the downside, the assumption of site independence makes the method vulnerable to violation of this assumption (Bustamante *et al.*, 2001). Zhu and Bustamante (2005) relaxed the assumption of site independence by incorporating recombination within a composite likelihood approach. Their composite likelihood ratio test showed good power to detect recurrent directional selection and was relatively robust to estimation bias of local recombination rate but not to population growth or a recent bottleneck. The power of Poisson random-field methods (as well as the composite likelihood method) can be increased by considering multiple loci, maximizing information about species divergence time and population sizes, which are shared among loci.

Tests based on the idea of between-species and within-species comparisons require population data as well as species sampling. This is not always feasible in macroevolutionary studies.

## Comparative genomic methods at the protein-coding DNA level

### The $d_N/d_S$ measure

The most direct way of obtaining evidence for adaptive evolution on a protein-coding gene is by comparing the nonsynonymous substitution rate  $d_N$  with the synonymous rate  $d_S$  (Yang and Bielawski, 2000). These rates are defined as numbers of nonsynonymous or synonymous substitutions per nonsynonymous or synonymous site, respectively. Selective pressure at the protein level is measured by the ratio  $\omega = d_N/d_S$ . If the amino acid change is deleterious, purifying selection reduces its fixation rate so that  $\omega < 1$ . When substitution has no effect on fitness, nonsynonymous substitutions occur at the same rate as synonymous ones and  $\omega = 1$ , suggesting neutral evolution. Only when amino acid changes offer a selective advantage are nonsynonymous changes fixed at a higher rate than synonymous changes and  $\omega > 1$  provides the evidence for recurrent diversifying selection.

The methods for estimating  $d_N$  and  $d_S$  used in early studies were so-called approximate methods, developed for pairwise calculations. Although they greatly differ in detail, the basic procedure is the same: count the numbers of synonymous and nonsynonymous sites in the observed sequences; count the numbers of synonymous and nonsynonymous differences by considering all possible evolutionary pathways between the homologous codons; and correct for multiple substitutions at the same sites by using a standard evolutionary model (Yang, 2006).

### Estimating $d_N/d_S$ by maximum likelihood

Since the development of approximate pairwise methods, a more accurate maximum likelihood method has gained in popularity (Yang and Nielsen, 2000). The maximum likelihood framework is convenient for hypotheses testing via likelihood ratio tests. The Bayesian prediction approach, easily adapted within a probabilistic framework, can be used to evaluate probable scenarios by quantifying the uncertainty of various predictions in an easily interpretable form. Although in some cases the maximum likelihood approach can be computationally intensive, approximate alternatives such as composite likelihood are possible (Fearnhead and Donnelly, 2001; Hudson, 2001; Zhu and Bustamante, 2005; Wilson and McVean, 2006).

The first models of codon evolution were formulated as a continuous-time Markov process of substitution along the phylogeny relating the sequences (Goldman and Yang, 1994; Muse and Gaut, 1994), and thus were applicable to multiple sequence alignments. Muse and Gaut incorporated only variable base frequencies and separate  $d_N$  and  $d_S$  rates. Goldman and Yang accounted for nonuniform codon frequencies and transition/transversion bias and used a separate parameter to measure gene variability (later simplified to include the  $\omega$  ratio explicitly; Nielsen and Yang, 1998).

Several assumptions are commonly made for computational convenience rather than to reflect biological reality. The substitution process is commonly assumed to be independent at each site, time reversible (as the direction of change in the observed data is unknown),

homogeneous (that is, the same throughout time) and stationary (that is, remains at the equilibrium throughout time; for example, codon frequencies are the same over the course of the evolution). Models relaxing such common assumptions were proposed (Galtier and Gouy, 1998; Galtier, 2001; Lartillot and Philippe, 2004; Pagel and Meade, 2004 among others; some will be discussed later).

### Partitioning of sites in $d_N/d_S$ estimation

Methods that assume a constant  $\omega$  ratio for all sites and over time detect positive selection only if the average  $\omega$  is greater than 1 (Yang and Bielawski, 2000). The first study that successfully detected positive selection by accounting for rate variability among sites partitioned residues in the Major Histocompatibility Complex class I into those from the antigen-recognition region and those outside this region, and compared  $d_N$  and  $d_S$  rates in each partition (Hughes and Nei, 1988). Yang and Swanson (2002) implemented a maximum likelihood method for pre-partitioned datasets, using a separate  $\omega$  for each partition. As *a priori* information on site functionality is not often available, most other methods for detecting positive selection do not require *a priori* knowledge. Several non-likelihood approaches partition sites in a sequence—for example, based on shifting conservation of sites under a covarion-like process (Siltberg and Liberles, 2002), by close linkage in gene structure (Fares *et al.*, 2002) or by using a tertiary windowing approach to lump together sites found in proximity in a protein's three-dimensional structure (Suzuki, 2004; Berglund *et al.*, 2005). Remaining methods may be subdivided into those that estimate  $\omega$  separately for each site (Fitch *et al.*, 1997; Suzuki and Gojobori, 1999; Nielsen and Huelsenbeck, 2002; Suzuki, 2004; Masingham and Goldman, 2005; Kosakovsky Pond and Frost, 2005b) and those that use a prior distribution to describe the selective pressure and use a Bayesian approach to classify sites (Nielsen and Yang, 1998; Yang *et al.*, 2000; Kosakovsky Pond and Muse, 2005).

### Ancestral sequences and $d_N/d_S$

Early non-likelihood methods accounting for variable selective pressures reconstruct sequences of the extinct ancestors; then, at each site, they count changes along the tree to identify sites with an excess of nonsynonymous substitutions (Fitch *et al.*, 1997; Messier and Stewart, 1997; Benner *et al.*, 1998; Bush *et al.*, 1999; Suzuki and Gojobori, 1999; Yamaguchi-Kabata and Gojobori, 2000; Liberles, 2001). These so-called 'counting' methods need large samples to ensure enough changes at codon sites. For divergent sequences (with long branch lengths in a corresponding phylogeny), the parsimonious ancestral reconstruction is unreliable. In particular, the inferred ancestral sequences would be less reliable at positively selected sites along long branches, as these are often the most variable sites in the alignment (Yang, 2006). Not only can the parsimonious solution be unlikely but nonparsimonious reconstructions can be much more likely (Nielsen, 2002). As parsimony focuses on reconstructions with minimum changes, it may seriously underestimate the total number of changes (Whelan and Goldman, 2001; Nielsen, 2002).

### Model testing and Bayesian site inference when $d_N/d_S$ varies among sites

As maximum likelihood models easily incorporate various molecular biases (explicitly as parameters) and avoid ancestral reconstruction, the existing codon models were adapted to allow variable selective pressure (Whelan and Goldman, 2001; Nielsen, 2002) and became extremely popular (MacCallum and Hill, 2006). Site codon models based on Goldman and Yang (1994) describe among-site variation in selective pressure by using a statistical distribution of  $\omega$  (implemented in PAML by Yang, 1997; Nielsen and Yang, 1998; Yang *et al.*, 2000); they can be used to test for positive selection and to detect affected sites by using Bayesian inference. To test for positive selection, a likelihood ratio test (LRT) is used to compare a null hypothesis that does not allow  $\omega > 1$  with a more general alternative hypothesis that allows  $\omega > 1$ . A gene is under positive selection if such an LRT is significant and one of the  $\omega$  estimates is  $> 1$  at a nonzero proportion of sites. Such LRTs were found to be accurate in simulations; their power correlates with data information content as measured by sequence divergence (Anisimova *et al.*, 2001).

If the LRT for positive selection is significant, sites under positive selection may be inferred with the Bayesian approach. The posterior probability that a site belongs to each  $\omega$  class of the model (given the data) provides an intuitive measure of confidence in a prediction. Sites with high posterior probabilities of belonging to a class with  $\omega > 1$  are likely to be under positive selection. For data with low information content (such as population data from slowly evolving closely related species), the posterior probabilities can be misleading (Anisimova *et al.*, 2002). The choice of parameter values describing the prior probabilities may influence the inference. Yang *et al.* (2000) used the naive empirical approach with maximum likelihood parameter estimates as priors. To deal with estimation uncertainty, Yang *et al.* (2005) implemented the Bayes empirical Bayes approach, in which only the selection-related parameters were numerically integrated over their prior distributions, whereas topology, branch lengths and codon frequencies remained fixed to best estimates. The full Bayesian solution obtained posterior probabilities by Markov Chain Monte Carlo (MCMC) approximation, integrating the conditional distribution over the assumed prior distribution for all nuisance parameters, including the tree and branch lengths (Huelsenbeck and Dyer, 2004). In simulations, the Bayes empirical Bayes approach outperformed the naive empirical approach and was largely as accurate as the full Bayesian approach (Scheffler and Seoighe, 2005; Aris-Brosou, 2006). Being computationally intensive, the full Bayesian approach is feasible only for small data sets; it may be beneficial for data that are either very similar or very divergent. Many cases of adaptation were detected with the site models of Yang *et al.* (2000) and were subsequently extensively studied in simulations (Anisimova *et al.*, 2001, 2002, 2003; Wong *et al.*, 2004; Massingham and Goldman, 2005; Scheffler and Seoighe, 2005).

LRTs for positive selection were found to be increasingly inaccurate with increases of recombination rate (Anisimova *et al.*, 2003). Two possible reasons for this inaccuracy are that the method relies on a single inferred phylogeny and that only the nonsynonymous rate

variation was incorporated in the model, whereas the synonymous rate was assumed to be constant. Consequently, Kosakovsky Pond and Muse (2005) incorporated both nonsynonymous and synonymous rate variation in their codon models (based on Muse and Gaut, 1994; implemented in HYPHY by Kosakovsky Pond *et al.*, 2005). Conceptually very similar to the site models of Yang *et al.* (2000), these models often produce nearly identical results. Accounting for  $d_S$  variation generally improves the fit of the model, yet the constancy of the synonymous rate may often be assumed without compromising accuracy. However, when certain aspects of the substitution process vary significantly among sites, failure to accommodate such variability can have a negative impact on maximum likelihood estimation (Bao *et al.*, 2007). In addition to  $d_S$  variation, Scheffler *et al.* (2006) allowed topology and branch lengths to vary across inferred recombination breakpoints, greatly improving the robustness of the test in simulations. When the method was applied to genes from human immunodeficiency virus (HIV)-1 with frequent recombination, positive selection reported in previous studies could not be confirmed. This led to a conclusion that positive selection on these genes was inferred falsely, as a result of reliance on a single phylogeny.

Unlike counting methods, the popular maximum likelihood models (Yang *et al.*, 2000; Kosakovsky Pond and Muse, 2005) make assumptions about the distribution of synonymous and nonsynonymous rates across sites. Other maximum likelihood methods estimate the  $\omega$  ratio for each site independently, making no assumptions about the underlying distribution of  $d_N$  and  $d_S$  (Huelsenbeck, 2002; Nielsen, 2002; Suzuki, 2004; Massingham and Goldman, 2005; Kosakovsky Pond and Frost, 2005b). Testing the null hypothesis ' $\omega = 1$ ' at each site, such methods may prove more robust to recombination if at each site they use a phylogeny consistent with the site history instead of relying on a single topology. Even when testing ' $\omega = 1$ ' at each site, short bursts of adaptive evolution may be difficult to detect while averaging over the history of a sample. Whether any sites in three-dimensional protein structures really evolve purely neutrally is unclear; this remains a subject for study, which affects the utility of this test.

In a novel approach, Wilson and McVean (2006) used a likelihood approximation to the coalescent process with recombination and used reversible-jump MCMC to perform Bayesian inference simultaneously on selection and recombination parameters, allowing each to vary among sites. Computer simulations showed that the method was accurate and had the power to detect positive selection in the presence of recombination.

Nielsen and Huelsenbeck (2002) proposed a counting-like Bayesian approach that avoids the problems caused by focusing on a single parsimonious ancestral reconstruction. Numbers of nonsynonymous and synonymous changes were formulated as functions of the mapping of mutations on a phylogeny (a reconstruction) and were inferred for any mapping directly. As the true mapping is unknown, all mapping possibilities are considered and weighted accordingly. Whereas the maximum likelihood method relies on the known phylogeny, Nielsen and Huelsenbeck treated it as a nuisance parameter, integrating it out through MCMC. In this framework, hypotheses are tested by comparing the posterior and posterior

predictive (expected) distributions of the statistic of interest (Meng, 1994). As with other Bayesian applications, the choice of prior distribution for the parameters of the model may influence the inference. Both Nielsen and Huelsenbeck (2002) and Yang *et al.* (2000) detected largely the same sites under positive selection in a subset of hemagglutinin (HA) flu data (Fitch *et al.*, 1997). In another attempt to use the advantages of likelihood, Suzuki (2004) used maximum likelihood to reconstruct ancestral amino acid sequences and parsimony to map codon states restricted by amino acid inferences. The  $\omega$  ratio was estimated for each codon with maximum likelihood, and an LRT was used to verify if  $\omega$  was significantly greater than 1. Kosakovsky Pond and Frost (2005b) implemented several counting maximum likelihood-based methods based on the ideas from Suzuki and Gojobori (1999) and Suzuki (2004); they used a likelihood-based analog of the site-by-site counting methods but estimated the  $\omega$  ratio for each site in a sequence alignment. A similar site-by-site maximum likelihood estimation method was proposed by Massingham and Goldman (2005). Kosakovsky Pond and Frost (2005b) explored several counting methods based on ancestral reconstruction, comparing these with methods assuming an *a priori* distribution of  $d_N$  and  $d_S$  rates and those estimating  $d_N$  and  $d_S$  at each site without relying on ancestral reconstruction. All tested methods were modifications of already existing techniques, in order to make fair comparisons and reconcile the performance differences among existing site methods. The conclusion 'not so different after all' confirmed that different approaches, when carefully implemented, generally led to almost identical results.

Even when  $d_N$  variation was modeled in a more flexible way through a Dirichlet process mixture model (Huelsenbeck *et al.*, 2006), the sites inferred under positive selection were in strong concordance with those inferred by Yang *et al.* (2000). The full Bayesian approach of Huelsenbeck *et al.* (2006) currently offers the most flexible description of  $\omega$  variation across sites while accounting for uncertainty in parameter estimates. However, the study also shows that increasing model complexity does not always contribute to more accurate inference.

#### Lineage-specific $d_N/d_S$ variation and the development of branch and branch-site codon models

The first to offer a way of detecting episodic positive selection on specific lineages was the study of primate lysozyme (Messier and Stewart, 1997) that reconstructed sequences of extinct ancestors in a primate phylogeny by using parsimony and maximum likelihood methods. Both reconstructed and observed sequences were used to estimate the average pairwise  $d_N$  and  $d_S$  rates for each branch of the tree. This analysis detected positive selection in two lineages: a lineage leading to the common ancestor of foregut fermenting colobine monkeys and a lineage leading to the common ancestor of the modern hominoid lysozymes. Crandall and Hillis (1997) took a similar approach based on maximum likelihood reconstruction to test the variability of selective constraints between the rhodopsin genes of cave-dwelling and surface-dwelling crayfishes. Although explicit reconstruction of ancestral sequences may be useful for

experimental testing of hypotheses in the laboratory, it may also introduce biases into the inference of positive selection, especially if single inferred ancestral sequences are treated as observed (Zhang *et al.*, 1997; Williams *et al.*, 2006; Pollock and Chang, 2007).

Yang (1998) implemented maximum likelihood codon models of independent  $\omega$  ratios for different branches of a tree, in which transition probabilities for different branches were calculated using instantaneous rate matrices with different  $\omega$  ratios. The most flexible lineage-specific model is the free-ratio model, which assumes a separate  $\omega$  parameter for each branch, whereas the simplest has the same  $\omega$  for all lineages. Intermediate models are constructed by specifying sets of branches with different  $\omega$  ratios. Yang (1998) used the lysozyme data to test hypotheses that made different assumptions about the  $\omega$  ratios on the branches reported under positive selection by Messier and Stewart, relative to the ratio on all the other branches. For example, a two-ratio model might assume that a branch ancestral to colobine monkeys has an  $\omega$  ratio different from others. An LRT is used to test whether a two-ratio model fits data significantly better than a one-ratio model. Recently, Kosakovsky Pond and Frost (2005a) developed a genetic algorithm to assign  $\omega$  ratios to lineages on a tree by 'evolving' the model to be tested through maximizing its fit. Such an approach is useful as *a priori* specification of lineages is not required, although the procedure does not offer a statistical test for positive selection (merely providing the estimates of selective pressure). Correction for multiple testing may be necessary if one is interested in a conservative set of candidate genes under positive selection with fewer false positives than false negatives (Yang, 2006; Anisimova and Yang, 2007).

All the mentioned branch methods assume a constant selection pressure among sites, and so have low power to detect episodic positive selection that has occurred at a few sites. Incorporating variability of selective constraints across sites and over time simultaneously was essential for successfully detecting episodic selection operating at a few sites. Yang and Nielsen (2002) proposed the first such models, in which selective pressure varied across sites, although at a subset of sites it also changed along a set of branches specified *a priori* (the foreground). Using an LRT, a branch-site model with positive selection on the foreground can be compared with a model that does not allow positive selection. As the initial branch-site tests (Yang and Nielsen, 2002) exhibited excessive false positives (Zhang, 2004), modifications were proposed (Yang *et al.*, 2005). Zhang *et al.* (2005) suggested the LRT based on modified models, with satisfactory accuracy and power.

Forsberg and Christiansen (2003) developed a model applicable if the branches of the phylogeny could be grouped *a priori* into two clades. An LRT compared the divergence in selective pressure between the clades. Bielawski and Yang (2004) implemented several clade models with two or three discrete  $\omega$  classes, inspired by the site models of Yang *et al.* (2000). One such model assumes three site classes. In two classes, the  $\omega$  ratio is constant in all lineages (a conserved site class with  $\omega < 1$  and a neutral class with  $\omega = 1$ ). In the third class, sites may evolve under different selective pressures in the two clades; their  $\omega$  ratios are not constrained. Estimates of

$\omega > 1$  for this class suggest positive selection due to clade-specific differences in selective pressure. Clade models are very effective for studying the dynamics of host-specific adaptations or the evolution of gene families by discovering the differences in selective regimes during the evolution of two paralogs following a duplication event (see Roth *et al.*, 2007 for a discussion of gene evolution after gene duplication).

One limitation of methods that require *a priori* specification of branches to be tested for positive selection (Yang and Nielsen, 2002; Forsberg and Christiansen, 2003; Bielawski and Yang, 2004; Yang *et al.*, 2005) is that a prespecified biological hypothesis may not always be available—for example, when the gene function is poorly understood or during an automatic genome-scale scan for positive selection. Anisimova and Yang (2007) suggested a possible approach whereby several or all branches on the tree are tested, with every branch treated in turn as a foreground branch. To avoid excessive false positives, a correction for multiple hypotheses testing has to be applied, such as the Benjamini and Hochberg (1995) procedure controlling the false-discovery rate. Despite the non-independence of tests, multiple tests correction was shown to be applicable in simulations, but exceptions may arise as a result of forces not considered in the study (Anisimova and Yang, 2007). For example, Johnston *et al.* (2007) suggest that positive selection on different branches in a gene family may not be independent. The application of corrections for multiple testing is still debated in the research community and deserves further study.

In the absence of an *a priori* hypothesis, the application of a multiple test correction makes the branch-sites models of Yang *et al.* overly conservative. An alternative branch-site approach generalized the site models with discrete classes (Yang *et al.*, 2000, 2005) to allow changes between the selection regimes (Guindon *et al.*, 2004). This approach is similar to the covarion model for site-specific variation of substitution rates, in which every site may switch between high and low rates (Tuffley and Steel, 1998; Galtier, 2001; Huelsenbeck, 2002). Zhai *et al.* (2007) extended the site approach of Nielsen and Huelsenbeck (2002) based on mutational mappings to include lineage-specific variation, but only a single inferred phylogeny was used to increase the computational efficiency. Likewise, the genetic algorithm branch approach (Kosakovsky Pond and Frost, 2005a) based on model selection can be extended to simultaneously accommodate site-to-site rate variation by adjusting site rates for the entire tree, but this becomes computationally expensive. A new codon-based serial model of evolution will permit changes to the selection intensities at sites (and the proportions of sites under different selective pressures) simultaneously across all lineages and will therefore be ideal for exploring the dynamics of disease progression after a specified time point (A Rodrigo, personal communication). This would be particularly relevant during environmental change that affects all lineages simultaneously or after the start of an antiretroviral therapy or other treatment. Methods allowing both spatial and temporal variation of selective pressures are in their infancy, and work remains to be done toward validating these new methods in simulations, as well as on real data.

### Comparative power of codon methods

Branch, site and branch-site tests may or may not detect positive selection, depending on the fraction of sites affected and the time during which selection was operating, as well as its strength. The shorter the adaptive episode, the more difficult it is to detect it with sites models, especially when selection acted on very few sites or a small percentage of branches, as seems to be the case in the reanalysis of selection reported in myostatin (Tellgren *et al.*, 2004; Pie and Alvares, 2006; Massey, MA, and DAL, manuscript in preparation). Equally, branch models may fail to detect positive selection on a branch as a result of averaging across sites when only a fraction of sites were affected by positive selection. Models allowing both spatial and temporal variation of selective pressure are expected to be more powerful than either site or branch tests performed separately (Yang *et al.*, 2005; Zhang *et al.*, 2005; Guindon *et al.*, 2004).

Studies focusing on properties of the protein changes may also want to consider using codon-based methods that evaluate the selective effects on physicochemical properties of amino acids rather than simply detecting excess nonsynonymous changes (Xia and Li, 1998; McClellan and McCracken, 2001; Sainudiin *et al.*, 2005; Wong *et al.*, 2006). An early version of this idea was the use of the PAM/ $d_S$  statistic in place of  $d_N/d_S$  (Liberles, 2001).

### Methods acting at the amino acid level

#### Selective pressure on $d_S$ and the need for amino acid-level models

Methods comparing  $d_N$  and  $d_S$  are not suitable for fast evolving genes from species of deep divergences, as silent sites become quickly saturated over time (Smith and Smith, 1996; Yang and Nielsen, 2000; Fares *et al.*, 2002). Moreover, the convenient positive selection criteria  $d_N/d_S > 1$  is applicable if synonymous substitutions are neutral, which may not always be the case (Chamary *et al.*, 2006). Any selective effects on  $d_S$  that are different from effects on  $d_N$  are of concern. Increasing evidence suggests that synonymous changes might affect splicing and mRNA stability (Chamary *et al.*, 2006; Parmley *et al.*, 2006). For example, codon bias received particular attention with respect to this problem. In mammals, codon bias is static—that is, intragenic codon usage is nonrandom but not significantly different across most species (Nakamura *et al.*, 2000). This can be thought of as stable selection on  $d_S$  that reduces the amount of mutation that becomes fixed through negative selection acting on synonymous sites (Liberles, 2001). However, codon usage may change in a lineage-specific manner, which will have the opposite effect of inflating  $d_S$  and may lead to an underestimation of  $\omega$ , also biasing it as a measure of selective pressure. No methods are currently available to deal with this problem at the codon level. Beyond traditional codon bias,  $d_S$  may actually affect protein folding, through a novel and poorly understood process (Kimchi-Sarfaty *et al.*, 2007). Whether such a phenomenon is common is unclear, but its effect on  $\omega$  may be similar to that of static codon bias, assuming that the process itself is constant across lineages. Finally, the existence of overlapping reading frames may additionally complicate the notion of the  $d_N/d_S$  ratio, as

synonymous substitutions in one frame may be constrained to avoid nonsynonymous changes in another reading frame. Early models dealing with measuring selection in overlapping coding regions have been developed using Hidden Markov Model methodology (McCauley and Hein, 2006; McCauley *et al.*, 2007).

#### Detecting functional divergence with amino acid-level models

The problems associated with codon data may be avoided by using approaches that consider the replacement amino acid rate alone. In a similar fashion to codon methods, amino acid-based methods search for site- or lineage-specific rate accelerations and residues subject to altered functional constraints. Such sites are likely to be contributing to the change in protein function over time. The heterogeneity of rates among sites is typically described as a  $\gamma$  distribution with an  $\alpha$ -shape parameter (Uzzel and Corbin, 1971; Yang, 1993; Gu *et al.*, 1995, 2001), or by using site-specific matrices (Bruno, 1996; Halpern and Bruno, 1998) and mixture models (Koshi and Goldstein, 1995, 1997; Goldman *et al.*, 1998; Lartillot and Philippe, 2004; Soyer and Goldstein, 2004). Using factors such as secondary structural unit and solvent accessibility, context-dependent substitution matrices were proposed (Koshi and Goldstein, 1996). Because evolutionary rates may depend not only on structural factors such as secondary structural unit or solvent accessibility, simpler fitness site-class models that instead define a substitution process dependent on the relative fitness of the amino acid in a particular position are an interesting development (Dimmic *et al.*, 2005).

One class of methods for detecting functional divergence, analogous to branch models, searches for a lineage-specific change in  $\alpha$  (Miyamoto and Fitch, 1995; Lockhart *et al.*, 1998; Penny *et al.*, 2001; Siltberg and Liberles, 2002). Other methods search for evidence of clade-specific rate shifts at individual sites (Lichtarge *et al.*, 1996; Armon *et al.*, 2001; Gu, 1999; Gaucher *et al.*, 2002; Pupko and Galtier, 2002; Blouin *et al.*, 2003; Landau *et al.*, 2005). This is analogous to branch-site codon models and similar in idea to covarion-like models (Galtier, 2001; Guindon *et al.*, 2004). For example, Gu (1999) proposed a simple stochastic model for estimating the degree of divergence between two prespecified clusters and testing its statistical significance, whereby a site-specific profile based on a hidden Markov model was used to identify amino acids responsible for these functional differences between two gene clusters. More flexible evolutionary models were incorporated in the maximum likelihood approach applicable to the simultaneous analysis of several gene clusters (Gu, 2001). This was extended (Gu, 2006) to evaluate site-specific shifts in amino acid properties, in comparison with site-specific rate shifts. Pupko and Galtier (2002) used the LRT to compare the maximum likelihood estimates of the replacement rate of an amino acid site in distinct subtrees. Illustrating the technique on mammalian mitochondrial protein sequences, they showed that the primate lineage reached its current adaptive landscape through episodes of positive selection at a few sites, enabling the fine-tuning of the three-dimensional protein structure to optimize its conserved function. The authors argued that adaptive change on the level of a protein

sequence may not necessarily correspond to an adaptive change in protein function but rather to the peaks in protein adaptive landscape reflecting the optimization of the protein function in a particular species or to long-term environmental changes. Galtier and Jean-Marie (2004) extended the covarion approach to allow switching between more than two classes of sites by using time-continuous space-discrete Markov-modulated Markov chains. In the maximum likelihood framework, Wang *et al.* (2007) combined features of two earlier models (Galtier, 2001; Huelsenbeck, 2002) into a general covarion model that allows evolutionary rates to switch not only between variable and invariable classes but also among different rates even when they are in a variable state. To validate the accuracy of functional shift methods, Abhiman and Sonnhammer (2005a, b) analyzed large datasets of proteins with known enzymatic functions and found that combining methods that detect sites conserved in two subfamilies and those with significantly different rates (using linear discriminant analysis) improved the accuracy of classification. As with codon tests, most tests for temporal amino acid rate variation assume *a priori* partitioning of sequences into groups and test for homogeneity of rate among the groups. In contrast, Dorman (2007) proposed a Bayesian method to infer significant shifts in selective pressure affecting many sites simultaneously without *a priori* specifying the branch expected to contain the divergence point. To demonstrate the power of the method, a divergence point separating two HIV subtypes was successfully detected between genetically distinct viral variants that have spread into different human populations with the AIDS epidemic. On the downside, the power is low for sequences of insufficient divergence and only shifts of considerable magnitude are detectable (Dorman, 2007).

Predicting a functional shift from sequence data alone can be useful for large-scale protein annotation (for example, databases FunShift by Abhiman and Sonnhammer, 2005a, and PhyloFacts by Krishnamurthy *et al.*, 2006). Most of the methods discussed so far can be used in an automated manner in a preliminary examination of the functional evolution of protein families. Considering the current wealth of molecular data, many protein families are large enough for an informative statistical analysis of substitution patterns produced by adaptation events. However, without knowledge of protein structure and the constraints imposed on each site, distinguishing neutral substitutions from those that substantially modify the function is difficult.

#### Incorporating protein structural constraints

All the methods discussed above assume independence of the evolution at sites—an unrealistic assumption given that structure dictates evolutionary interdependence between protein residues. A default structural scenario is to analyze lineage-specific change in a structural context, using a force field to examine energetic effects of substitutions. Consequently, interest has increased in developing integrated models with general interdependence between the protein residues through explicitly incorporating structural constraints within an evolutionary phylogenetic framework. Such approaches require a known tertiary structure of a reference protein and involve measuring composition of site dependencies by

using statistical potentials—an empirical energy function relating the pseudoenergy terms to the plausibility of spatial proximity for a given residue pair, derived in a context of protein threading. This can be done by using a simple force field (Parisi and Echave, 2001; Rastogi *et al.*, 2006) or pairwise interaction matrices (Miyazawa and Jernigan, 1985; Jones *et al.*, 1992; Bastolla *et al.*, 2001). For example, Parisi and Echave (2001) used a protein evolution simulation that proposes amino acid replacements dependent on statistical potentials and discards sequences resulting in structurally divergent proteins. This was extended to the use of several different informational and force field methods by Rastogi *et al.* (2006). Fornasari *et al.* (2002) exploited the original simulation procedure to build replacement matrices incorporated into a phylogenetic context, which led to improved model fit (Parisi and Echave, 2004, 2005). Later studies incorporated a set of statistical potentials directly within a Markov substitution process (Robinson *et al.*, 2003; Rodrigue *et al.*, 2005, 2006). Using Bayesian model selection methods, alternative ways to explicitly model structural constraints (or other site-interdependencies) can be explored (Rodrigue *et al.*, 2006). All such models assume a conservation of protein structure over the phylogenetic history of a sample, an assumption that is unlikely to be problematic in most cases. Other models incorporate site interdependence without explicitly using structural information (Fares and Travers, 2006; Stern and Pupko, 2006; Kalinina *et al.*, 2007). Extension of structural phylogenetic methods to the explicit detection of functional change and the relationship between thermodynamics and selection will be important future directions.

## Selecting adequate methods for comparative genomic studies

It is worth looking back at several large-scale studies to learn about the advantages and the drawbacks of the methods used. In a landmark large-scale search for positive selection, Endo *et al.* (1996) estimated pairwise  $\omega$  ratios for 3595 genes, confirming positive selection for only 0.45% of analyzed genes. Other large-scale scans for positive selection using a non-likelihood branches approach found a several-fold higher incidence of positive selection (Liberles *et al.*, 2001; Roth *et al.*, 2005; Roth and Liberles, 2006). Although neither applied a sites approach, the later studies had better power to detect positive selection as their approach used information from more than two lineages for each gene. Successful detection of positive selection, while averaging across sites, may serve as evidence that sometimes evolution does act in a punctuated manner.

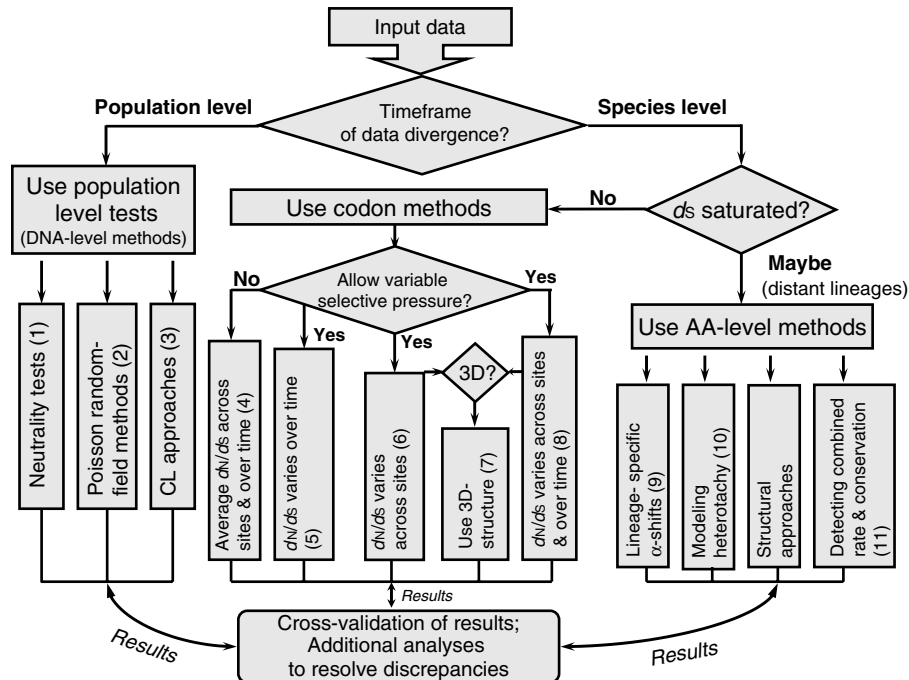
Maximum likelihood and Bayesian methods were used to test for adaptive changes and to estimate the strength of selection in population data (Bustamante *et al.*, 2002, 2003). Bustamante *et al.* (2002) concluded that in *Drosophila* substitutions were predominantly beneficial, whereas in *Arabidopsis* they were predominantly deleterious. The difference was attributed to partial self-mating in *Arabidopsis*, making it difficult for the species to eliminate deleterious mutations.

Many studies evaluating and modifying site and branch-site LRTs (Anisimova *et al.*, 2001, 2002, 2003;

Wong *et al.*, 2004; Yang *et al.*, 2005; Zhang *et al.*, 2005; Anisimova and Yang, 2007) demonstrate the use of these methods in large-scale studies but also point out the conditions causing them to become inaccurate and warn against overgeneralizing. These methods and the population genetics approaches discussed above complemented each other in inferences of adaptation in model organisms. For example, the sources of human–chimp divergence became a hot subject in comparative genomics (Clark *et al.*, 2003; Bustamante *et al.*, 2005; Nielsen *et al.*, 2005; Arbiza *et al.*, 2006). The combination of methods used produced some disagreement among evolutionary biologists (Arbiza *et al.*, 2006); nevertheless, the studies provided some important biological insights into the nature of selective pressure during the divergence of humans from their ancestors. Clark *et al.* (2003) analyzed orthologous human–chimp–mouse trios with a branch-site test comparing the neutral codon model with a more general model that allowed sites with  $\omega > 1$  in the human lineage. Genes with accelerated evolution in the human lineage included those involved in sensory perception, amino acid catabolism and nuclear transport. Continuing the search for human genes under positive selection, Nielsen *et al.* (2005) used maximum likelihood pairwise lineage comparisons across a larger set of genes, including all human–chimp orthologs. Genes displaying evidence for positive selection included immune defense-related and sensory perception genes, as well as those involved in spermatogenesis, tumor suppression and apoptosis. An analysis of human polymorphism data from genes with the strongest signal of positive selection showed an excess of high-frequency derived nonsynonymous mutations, confirming the signal of positive selection in these genes. Bustamante *et al.* (2005) used a comparative population genomics technique (Sawyer and Hartl, 1992; Bustamante *et al.*, 2002; Sawyer *et al.*, 2003) to compare DNA polymorphism within humans with the human–chimp interspecies divergence, identifying a partially overlapping list of genes under positive natural selection along the human lineage. The test is subject to different assumptions and also has power over a different timescale. In the latest comparative genomic study of humans, chimps and their mural ancestors, Arbiza *et al.* (2006) used a more accurate branch test (Zhang *et al.*, 2005), aiming to differentiate the positive selection events from those that may have been caused by relaxation of selection constraints. Recently, Ardawatia and Liberles (2007) extended the area of human–chimp lineage comparison to include a systematic analysis of lineage-specific evolution in metabolic pathways by consolidating positive selection inferences with Kyoto Encyclopedia of Genes and Genomes pathway data (Kanehisa *et al.*, 2004).

## Conclusion

No recipe for a perfect comparative genomics analysis exists, but using a variety of techniques as illustrated above contributes to the gradual unraveling of the core evolutionary mechanisms. The flowchart in Figure 1 summarizes the main categories of available techniques. Essentially, the success in detecting positive selection depends on the strength of the signal and other evolutionary forces acting on the sequence in combination with the use of an appropriate method. For example,



**Figure 1** Flowchart representing a simple choice strategy for selection analyses. The timescale of sequence divergence dictates the types of appropriate techniques. Complexity of the models used depends on the scale of the study: large-scale and whole genome analyses are limited by the choice of computationally efficient techniques that do not require *a priori* knowledge. Analyses of individual genes or gene families may enjoy a greater choice of models, incorporating greater complexity or requiring *a priori* knowledge or resolved protein structure. References to methods indicated in the flowchart and mentioned in the text are listed below: (1) Tajima, 1989; Fu and Li, 1993; Fay and Wu, 2000; Hudson *et al.*, 1987; McDonald and Kreitman, 1991; Akashi, 1995, 1999b; Templeton, 1996; Eyre-Walker, 2002; Fay *et al.*, 2001, 2002; Smith and Eyre-Walker, 2002. (2) Hartl *et al.*, 1994; Sawyer and Hartl, 1992; Akashi, 1999a. (3) Zhu and Bustamante, 2005; Wilson and McVean, 2006. (4) Yang and Nielsen, 2000; Goldman and Yang, 1994; Muse and Gaut, 1994; for review of approximate methods see Yang and Bielawski (2000) and Yang (2006). (5) Messier and Stewart, 1997; Crandall and Hillis, 1997; Zhang *et al.* 1997; Williams *et al.*, 2006; Pollock and Chang, 2007; Yang, 1998; Kosakovsky Pond and Frost, 2005a. (6) Hughes and Nei, 1988; Fitch *et al.*, 1997; Bush *et al.*, 1999; Suzuki and Gojobori, 1999; Yamaguchi-Kabata and Gojobori, 2000; Yang and Swanson, 2002; Siltberg and Liberles, 2002; Fares *et al.*, 2002; Kosakovsky Pond and Frost, 2005b; Masingham and Goldman, 2005; Nielsen and Huelsenbeck, 2002; Kosakovsky Pond and Muse, 2005; Nielsen and Yang, 1998; Yang *et al.*, 2000; Huelsenbeck and Dyer, 2004; Bao *et al.*, 2007; Scheffler *et al.*, 2006; Wilson and McVean, 2006; Huelsenbeck *et al.*, 2006. (7) Suzuki, 2004; Berglund *et al.*, 2005. (8) Yang and Nielsen, 2002; Yang *et al.*, 2005; Zhang *et al.*, 2005; Forsberg and Christiansen, 2003; Bielawski and Yang, 2004; Anisimova and Yang, 2007; Guindon *et al.*, 2004; Zhai *et al.*, 2007. (9) Miyamoto and Fitch, 1995; Lockhart *et al.*, 1998; Penny *et al.*, 2001; Siltberg and Liberles, 2002. (10) Lichtarge *et al.*, 1996; Gu, 1999, 2001, 2006; Armon *et al.*, 2001; Gaucher *et al.*, 2002; Pupko and Galtier, 2002; Blouin *et al.*, 2003; Landau *et al.*, 2005; Galtier and Jean-Marie, 2004; Wang *et al.*, 2007; Dorman, 2007. (11) Abhiman and Sonnhammer, 2005b.

different methods will detect selective sweeps versus direct selection on a binding pocket of a protein. Understanding the applicability conditions for methods used is clearly essential. Interpretation of the results is equally important. For example, an excess of amino acid substitutions may not always be sufficient to prove adaptive evolution, as it can result from selection on synonymous substitutions rather than positive selection on a protein (Chamary *et al.*, 2006). To verify the consistency of results and their interpretation, use of several methods making different assumptions is good practice. Absence of evidence for positive selection with one method does not imply its nonexistence or that another method may not detect it. Equally, inferred cases of positive selection are strengthened if they are reconfirmed with diverse statistical and structural approaches as well as experimental studies testing for functional change. Ultimately, a fuller picture of the natural forces generating the patterns observed at the molecular level will emerge.

The era of genomics has proceeded hand in hand with the development of novel large-scale tests for selection. The age of comparative genomics is expected to extend

into the age of population genomics. This will enable the application of statistical tests applied across populations (population-level tests), between closely related species (DNA-level tests for protein encoding genes, where  $d_S$  has not reached saturation) and in analysis of more anciently diverged species (amino acid-level tests). Through these various comparisons, we will gain understanding of the interplay between population genetics, genomic forces, molecular and cellular constraints and thermodynamics at the protein structural level, coupled to lineage-specific adaptation to drive sequence evolution and species divergence.

## Acknowledgements

We thank Nicolas Galtier, Joe Bielawski and anonymous reviewers for carefully reading the manuscript and for their comments. MA was supported by a BBSRC (UK) grant.

## References

- Abhiman S, Sonnhammer EL (2005a). FunShift: a database of function shift analysis on protein subfamilies. *Nucleic Acids Res* 33: D197–D200.

- Abhiman S, Sonnhammer EL (2005b). Large-scale prediction of function shift in protein families with a focus on enzymatic function. *Proteins* **60**: 758–768.
- Akashi H (1995). Inferring weak selection from patterns of polymorphism and divergence at ‘silent’ sites in *Drosophila* DNA. *Genetics* **139**: 1067–1076.
- Akashi H (1999a). Inferring the fitness effects of DNA mutations from polymorphism and divergence data: statistical power to detect directional selection under stationarity and free recombination. *Genetics* **151**: 221–238.
- Akashi H (1999b). Within- and between-species DNA sequence variation and the ‘footprint’ of natural selection. *Gene* **238**: 39–51.
- Anisimova M, Bielawski JP, Yang Z (2001). Accuracy and power of the likelihood ratio test in detecting adaptive molecular evolution. *Mol Biol Evol* **18**: 1585–1592.
- Anisimova M, Bielawski JP, Yang Z (2002). Accuracy and power of bayes prediction of amino acid sites under positive selection. *Mol Biol Evol* **19**: 950–958.
- Anisimova M, Nielsen R, Yang Z (2003). Effect of recombination on the accuracy of the likelihood method for detecting positive selection at amino acid sites. *Genetics* **164**: 1229–1236.
- Anisimova M, Yang Z (2007). Multiple hypothesis testing to detect lineages under positive selection that affects only a few sites. *Mol Biol Evol* **24**: 1219–1228.
- Arbiza L, Dopazo J, Dopazo H (2006). Positive selection, relaxation, and acceleration in the evolution of the human and chimp genome. *PLoS Comput Biol* **2**: e38.
- Ardawatia H, Liberles DA (2007). A systematic analysis of lineage-specific evolution in metabolic pathways. *Gene* **387**: 67–74.
- Aris-Brosou S (2006). Identifying sites under positive selection with uncertain parameter estimates. *Genome* **49**: 767–776.
- Armon A, Graur D, Ben-Tal N (2001). ConSurf: an algorithmic tool for the identification of functional regions in proteins by surface mapping of phylogenetic information. *J Mol Biol* **307**: 447–463.
- Bao L, Gu H, Dunn KA, Bielawski JP (2007). Methods for selecting fixed-effect models for heterogeneous codon evolution, with comments on their application to gene and genome data. *BMC Evol Biol* **7** (Suppl 1): S5.
- Bastolla U, Farwer J, Knapp EW, Vendruscolo M (2001). How to guarantee optimal stability for most representative structures in the protein data bank. *Proteins* **44**: 79–96.
- Benjamini Y, Hochberg Y (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc B (Methodological)* **57**: 289–300.
- Benner SA, Sassi SO, Gaucher EA (2007). Molecular paleoscience: systems biology from the past. *Adv Enzymol Relat Areas Mol Biol* **75**: 1–132 xi.
- Benner SA, Trabesinger N, Schreiber D (1998). Post-genomic science: converting primary structure into physiological function. *Adv Enzyme Regul* **38**: 155–180.
- Berglund AC, Wallner B, Elofsson A, Liberles DA (2005). Tertiary windowing to detect positive diversifying selection. *J Mol Evol* **60**: 499–504.
- Bielawski JP, Yang Z (2004). A maximum likelihood method for detecting functional divergence at individual codon sites, with application to gene family evolution. *J Mol Evol* **59**: 121–132.
- Bloom JD, Raval A, Wilke CO (2007). Thermodynamics of neutral protein evolution. *Genetics* **175**: 255–266.
- Blouin C, Boucher Y, Roger AJ (2003). Inferring functional constraints and divergence in protein families using 3D mapping of phylogenetic information. *Nucleic Acids Res* **31**: 790–797.
- Bruno WJ (1996). Modeling residue usage in aligned protein sequences via maximum likelihood. *Mol Biol Evol* **13**: 1368.
- Bush RM, Fitch WM, Bender CA, Cox NJ (1999). Positive selection on the H3 hemagglutinin gene of human influenza virus A. *Mol Biol Evol* **16**: 1457–1465.
- Bustamante CD, Fledel-Alon A, Williamson S, Nielsen R, Hubisz MT, Glanowski S *et al.* (2005). Natural selection on protein-coding genes in the human genome. *Nature* **437**: 1153–1157.
- Bustamante CD, Nielsen R, Hartl DL (2003). Maximum likelihood and Bayesian methods for estimating the distribution of selective effects among classes of mutations using DNA polymorphism data. *Theor Popul Biol* **63**: 91–103.
- Bustamante CD, Nielsen R, Sawyer SA, Olsen KM, Purugganan MD, Hartl DL (2002). The cost of inbreeding in *Arabidopsis*. *Nature* **416**: 531–534.
- Bustamante CD, Wakeley J, Sawyer S, Hartl DL (2001). Directional selection and the site-frequency spectrum. *Genetics* **159**: 1779–1788.
- Chamary JV, Parmley JL, Hurst LD (2006). Hearing silence: non-neutral evolution at synonymous sites in mammals. *Nat Rev Genet* **7**: 98–108.
- Clark AG, Glanowski S, Nielsen R, Thomas PD, Kejariwal A, Todd MA *et al.* (2003). Inferring nonneutral evolution from human-chimp-mouse orthologous gene trios. *Science* **302**: 1960–1963.
- Crandall KA, Hillis DM (1997). Rhodopsin evolution in the dark. *Nature* **387**: 667–668.
- Depristo MA, Weinreich DM, Hartl DL (2005). Missense meanderings in sequence space: a biophysical view of protein evolution. *Nat Rev Genet* **6**: 678–687.
- Dimmic MW, Hubisz MJ, Bustamante CD, Nielsen R (2005). Detecting coevolving amino acid sites using Bayesian mutational mapping. *Bioinformatics* **21** (Suppl 1): i126–i135.
- Dorman KS (2007). Identifying dramatic selection shifts in phylogenetic trees. *BMC Evol Biol* **7** (Suppl 1): S10.
- Dutheil J, Pupko T, Jean-Marie A, Galtier N (2005). A model-based approach for detecting coevolving positions in a molecule. *Mol Biol Evol* **22**: 1919–1928.
- Endo T, Ikeo K, Gojobori T (1996). Large-scale search for genes on which positive selection may operate. *Mol Biol Evol* **13**: 685–690.
- Eyre-Walker A (2002). Changing effective population size and the McDonald–Kreitman test. *Genetics* **162**: 2017–2024.
- Fares MA, Elena SF, Ortiz J, Moya A, Barrio E (2002). A sliding window-based method to detect selective constraints in protein-coding genes and its application to RNA viruses. *J Mol Evol* **55**: 509–521.
- Fares MA, Travers SA (2006). A novel method for detecting intramolecular coevolution: adding a further dimension to selective constraints analyses. *Genetics* **173**: 9–23.
- Fay JC, Wu CI (2000). Hitchhiking under positive Darwinian selection. *Genetics* **155**: 1405–1413.
- Fay JC, Wyckoff GJ, Wu CI (2001). Positive and negative selection on the human genome. *Genetics* **158**: 1227–1234.
- Fay JC, Wyckoff GJ, Wu CI (2002). Testing the neutral theory of molecular evolution with genomic data from *Drosophila*. *Nature* **415**: 1024–1026.
- Fearnhead P, Donnelly P (2001). Estimating recombination rates from population genetic data. *Genetics* **159**: 1299–1318.
- Fitch WM, Bush RM, Bender CA, Cox NJ (1997). Long term trends in the evolution of H(3) HA1 human influenza type A. *Proc Natl Acad Sci USA* **94**: 7712–7718.
- Fornasari MS, Parisi G, Echave J (2002). Site-specific amino-acid replacement matrices from structurally constrained protein evolution simulations. *Mol Biol Evol* **19**: 352–356.
- Forsberg R, Christiansen FB (2003). A codon-based model of host-specific selection in parasites, with an application to the influenza A virus. *Mol Biol Evol* **20**: 1252–1259.
- Fu YX, Li WH (1993). Statistical tests of neutrality of mutations. *Genetics* **133**: 693–709.
- Galtier N (2001). Maximum-likelihood phylogenetic analysis under a covarion-like model. *Mol Biol Evol* **18**: 866–873.
- Galtier N, Gouy M (1998). Inferring pattern and process: maximum-likelihood implementation of a nonhomogeneous model of DNA sequence evolution for phylogenetic analysis. *Mol Biol Evol* **15**: 871–879.

- Galtier N, Jean-Marie A (2004). Markov-modulated Markov chains and the covarion process of molecular evolution. *J Comput Biol* **11**: 727–733.
- Gaucher EA, Gu X, Miyamoto MM, Benner SA (2002). Predicting functional divergence in protein evolution by site-specific rate shifts. *Trends Biochem Sci* **27**: 315–321.
- Golding GB, Dean AM (1998). The structural basis of molecular adaptation. *Mol Biol Evol* **15**: 355–369.
- Goldman N, Thorne JL, Jones DT (1998). Assessing the impact of secondary structure and solvent accessibility on protein evolution. *Genetics* **149**: 445–448.
- Goldman N, Yang Z (1994). A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol Biol Evol* **11**: 725–736.
- Gould SJ, Eldredge N (1993). Punctuated equilibrium comes of age. *Nature* **366**: 223–227.
- Gu X (1999). Statistical methods for testing functional divergence after gene duplication. *Mol Biol Evol* **16**: 1664–1674.
- Gu X (2001). Maximum-likelihood approach for gene family evolution under functional divergence. *Mol Biol Evol* **18**: 453–464.
- Gu X (2006). A simple statistical method for estimating type-II (cluster-specific) functional divergence of protein sequences. *Mol Biol Evol* **23**: 1937–1945.
- Gu X, Fu YX, Li WH (1995). Maximum likelihood estimation of the heterogeneity of substitution rate among nucleotide sites. *Mol Biol Evol* **12**: 546–557.
- Gu X, Zheng Y, Huong H, Xu D (2007). Using ancestral sequence inference to determine the trend of functional divergence after gene duplication. In: Liberles DA (ed). *Ancestral Sequence Reconstruction*. Oxford University Press: Oxford.
- Guindon S, Rodrigo AG, Dyer KA, Huelsenbeck JP (2004). Modeling the site-specific variation of selection patterns along lineages. *Proc Natl Acad Sci USA* **101**: 12957–12962.
- Haag ES (2007). Compensatory vs. pseudocompensatory evolution in molecular and developmental interactions. *Genetica* **129**: 45–55.
- Halpern AL, Bruno WJ (1998). Evolutionary distances for protein-coding sequences: modeling site-specific residue frequencies. *Mol Biol Evol* **15**: 910–917.
- Hartl DL, Moriyama EN, Sawyer SA (1994). Selection intensity for codon bias. *Genetics* **138**: 227–234.
- Hinman VF, Nguyen AT, Cameron RA, Davidson EH (2003). Developmental gene regulatory network architecture across 500 million years of echinoderm evolution. *Proc Natl Acad Sci USA* **100**: 13356–13361.
- Hudson RR (2001). Two-locus sampling distributions and their application. *Genetics* **159**: 1805–1817.
- Hudson RR, Kreitman M, Aguade M (1987). A test of neutral molecular evolution based on nucleotide data. *Genetics* **116**: 153–159.
- Huelsenbeck JP (2002). Testing a covarion model of DNA substitution. *Mol Biol Evol* **19**: 698–707.
- Huelsenbeck JP, Dyer KA (2004). Bayesian estimation of positively selected sites. *J Mol Evol* **58**: 661–672.
- Huelsenbeck JP, Jain S, Frost SW, Kosakovsky Pond SL (2006). A Dirichlet process model for detecting positive selection in protein-coding DNA sequences. *Proc Natl Acad Sci USA* **103**: 6263–6268.
- Hughes AL, Nei M (1988). Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. *Nature* **335**: 167–170.
- Johnston CR, O'Dushlaine C, Fitzpatrick DA, Edwards RJ, Shields DC (2007). Evaluation of whether accelerated protein evolution in chordates has occurred before, after, or simultaneously with gene duplication. *Mol Biol Evol* **24**: 315–323.
- Jones DT, Taylor WR, Thornton JM (1992). A new approach to protein fold recognition. *Nature* **358**: 86–89.
- Kalinina OV, Russel RB, Rakhmaninova AB, Gelfand MS (2007). Computational method for predicting protein functional sites with the use of specificity determinants. *Mol Biology* **41**: 137–147.
- Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M (2004). The KEGG resource for deciphering the genome. *Nucleic Acids Res* **32**: D277–D280.
- Kimchi-Sarfaty C, Oh JM, Kim IW, Sauna ZE, Calcagno AM, Ambudkar SV *et al.* (2007). A 'silent' polymorphism in the MDR1 gene changes substrate specificity. *Science* **315**: 525–528.
- King MC, Wilson AC (1975). Evolution at two levels in humans and chimpanzees. *Science* **188**: 107–116.
- Kosakovsky Pond SL, Frost SD (2005a). A genetic algorithm approach to detecting lineage-specific variation in selection pressure. *Mol Biol Evol* **22**: 478–485.
- Kosakovsky Pond SL, Frost SD (2005b). Not so different after all: a comparison of methods for detecting amino acid sites under selection. *Mol Biol Evol* **22**: 1208–1222.
- Kosakovsky Pond SL, Frost SD, Muse SV (2005). HyPhy: hypothesis testing using phylogenies. *Bioinformatics* **21**: 676–679.
- Kosakovsky Pond SL, Muse SV (2005). Site-to-site variation of synonymous substitution rates. *Mol Biol Evol* **22**: 2375–2385.
- Koshi JM, Goldstein RA (1995). Context-dependent optimal substitution matrices. *Protein Eng* **8**: 641–645.
- Koshi JM, Goldstein RA (1996). Probabilistic reconstruction of ancestral protein sequences. *J Mol Evol* **42**: 313–320.
- Koshi JM, Goldstein RA (1997). Mutation matrices and physical-chemical properties: correlations and implications. *Proteins* **27**: 336–344.
- Krishnamurthy N, Brown DP, Kirshner D, Sjolander K (2006). PhyloFacts: an online structural phylogenomic encyclopedia for protein functional and structural classification. *Genome Biol* **7**: R83.
- Landau M, Mayrose I, Rosenberg Y, Glaser F, Martz E, Pupko T *et al.* (2005). ConSurf 2005: the projection of evolutionary conservation scores of residues on protein structures. *Nucleic Acids Res* **33**: W299–W302.
- Lartillot N, Philippe H (2004). A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Mol Biol Evol* **21**: 1095–1109.
- Liberles DA (2001). Evaluation of methods for determination of a reconstructed history of gene sequence evolution. *Mol Biol Evol* **18**: 2040–2047.
- Liberles DA, Schreiber DR, Govindarajan S, Chamberlin SG, Benner SA (2001). The adaptive evolution database (TAED). *Genome Biol* **2**: R28.
- Lichtarge O, Bourne HR, Cohen FE (1996). An evolutionary trace method defines binding surfaces common to protein families. *J Mol Evol* **257**: 342–358.
- Lin YS, Hsu WL, Hwang JK, Li W-H (2007). Proportion of solvent-exposed amino-acids in a protein and rate of protein evolution. *Mol Biol Evol* **24**: 1005–1111.
- Lockhart PJ, Steel MA, Barbrook AC, Huson DH, Charleston MA, Howe CJ (1998). A covarion model explains apparent phylogenetic structure of oxygenic photosynthetic lineages. *Mol Biol Evol* **15**: 1183–1188.
- Lopez P, Casane D, Philippe H (2002). Heterotachy, an important process of protein evolution. *Mol Biol Evol* **19**: 1–7.
- MacCallum C, Hill E (2006). Being positive about selection. *PLoS Biol* **4**: e87.
- Massingham T, Goldman N (2005). Detecting amino acid sites under positive selection and purifying selection. *Genetics* **169**: 1753–1762.
- McCauley S, de Groot S, Hailund T, Hein J (2007). Annotation of selection strengths in viral genomes. *Bioinformatics* (in press).
- McCauley S, Hein J (2006). Using hidden Markov models and observed evolution to annotate viral genomes. *Bioinformatics* **22**: 1308–1316.
- McClellan DA, McCracken KG (2001). Estimating the influence of selection on the variable amino acid sites of the cytochrome B protein functional domains. *Mol Biol Evol* **18**: 917–925.

- McDonald JH, Kreitman M (1991). Adaptive protein evolution at the Adh locus in *Drosophila*. *Nature* **351**: 652–654.
- Meng X-L (1994). Posterior predictive *P*-values. *Ann Stat* **22**: 1142–1160.
- Messier W, Stewart CB (1997). Episodic adaptive evolution of primate lysozymes. *Nature* **385**: 151–154.
- Miyamoto MM, Fitch WM (1995). Testing the covarion hypothesis of molecular evolution. *Mol Biol Evol* **12**: 503–513.
- Miyazawa S, Jernigan RL (1985). Estimation of effective inter-residue contact energies from protein crystal structures—quasi-chemical approximation. *Macromolecules* **18**: 534–552.
- Muse SV, Gaut BS (1994). A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. *Mol Biol Evol* **11**: 715–724.
- Nakamura Y, Gojoberi T, Ikemura T (2000). Codon usage tabulated from international DNA sequence databases: status for the year 2000. *Nucleic Acids Res* **28**: 292.
- Nielsen R (2001). Statistical tests of selective neutrality in the age of genomics. *Heredity* **86**: 641–647.
- Nielsen R (2002). Mapping mutations on phylogenies. *Syst Biol* **51**: 729–739.
- Nielsen R, Bustamante C, Clark AG, Glanowski S, Sackton TB, Hubisz MJ *et al.* (2005). A scan for positively selected genes in the genomes of humans and chimpanzees. *PLoS Biol* **3**: e170.
- Nielsen R, Huelsenbeck JP (2002). Detecting positively selected amino acid sites using posterior predictive *P*-values. *Pacific Symposium on Biocomputing*, pp 576–588.
- Nielsen R, Yang Z (1998). Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. *Genetics* **148**: 929–936.
- Page M, Meade A (2004). A phylogenetic mixture model for detecting pattern-heterogeneity in gene sequence or character-state data. *Syst Biol* **53**: 571–581.
- Parisi G, Echave J (2001). Structural constraints and emergence of sequence patterns in protein evolution. *Mol Biol Evol* **18**: 750–756.
- Parisi G, Echave J (2004). The structurally constrained protein evolution model accounts for sequence patterns of the LbetaH superfamily. *BMC Evol Biol* **4**: 41.
- Parisi G, Echave J (2005). Generality of the structurally constrained protein evolution model: assessment on representatives of the four main fold classes. *Gene* **345**: 45–53.
- Parmley JL, Chamary JV, Hurst LD (2006). Evidence for purifying selection against synonymous mutations in mammalian exonic splicing enhancers. *Mol Biol Evol* **23**: 301–309.
- Penny D, McComish BJ, Charleston MA, Hendy MD (2001). Mathematical elegance with biochemical realism: the covarion model of molecular evolution. *J Mol Evol* **53**: 711–753.
- Pie MR, Alvares LE (2006). Evolution of myostatin in vertebrates: is there evidence for positive selection? *Mol Phylogenet Evol* **41**: 730–734.
- Pollock DD, Chang BSW (2007). Dealing with uncertainty in ancestral sequence reconstruction: sampling from the posterior distribution. In: Liberles DA (ed). *Ancestral Sequence Reconstruction*. Oxford University Press: Oxford.
- Pollock DD, Taylor WR, Goldman N (1999). Coevolving protein residues: maximum likelihood identification and relationship to structure. *J Mol Biol* **287**: 187–198.
- Prud'homme B, Gompel N, Rokas A, Kassner VA, Williams TM, Yeh SD *et al.* (2006). Repeated morphological evolution through cis-regulatory changes in a pleiotropic gene. *Nature* **440**: 1050–1053.
- Pupko T, Galtier N (2002). A covarion-based method for detecting molecular adaptation: application to the evolution of primate mitochondrial genomes. *Proc Biol Sci* **269**: 1313–1316.
- Rastogi S, Reuter N, Liberles DA (2006). Evaluation of models for the evolution of protein sequences and functions under structural constraint. *Biophys Chem* **124**: 134–144.
- Robinson DM, Jones DT, Kishino H, Goldman N, Thorne JL (2003). Protein evolution with dependence among codons due to tertiary structure. *Mol Biol Evol* **20**: 1692–1704.
- Rodrigue N, Lartillot N, Bryant D, Philippe H (2005). Site interdependence attributed to tertiary structure in amino acid sequence evolution. *Gene* **347**: 207–217.
- Rodrigue N, Philippe H, Lartillot N (2006). Assessing site-interdependent phylogenetic models of sequence evolution. *Mol Biol Evol* **23**: 1762–1775.
- Roth C, Betts MJ, Steffansson P, Saelensminde G, Liberles DA (2005). The adaptive evolution data base (TAED): a phylogeny based tool for comparative genomics. *Nucleic Acids Res* **33**: D495–D497.
- Roth C, Liberles DA (2006). A systematic search for positive selection in higher plants (Embryophytes). *BMC Plant Biol* **6**: 12.
- Roth C, Rastogi S, Arvestad L, Dittmar K, Light S, Ekman D *et al.* (2007). Evolution after gene duplication: models, mechanisms, sequences, systems, and organisms. *J Exp Zool B Mol Dev Evol* **308**: 58–73.
- Sainudiin R, Wong WS, Yogeewaran K, Nasrallah JB, Yang Z, Nielsen R (2005). Detecting site-specific physicochemical selective pressures: applications to the class I HLA of the human major histocompatibility complex and the SRK of the plant sporophytic self-incompatibility system. *J Mol Evol* **60**: 315–326.
- Sawyer SA, Hartl DL (1992). Population genetics of polymorphism and divergence. *Genetics* **132**: 1161–1176.
- Sawyer SA, Kulathinal RJ, Bustamante CD, Hartl DL (2003). Bayesian analysis suggests that most amino acid replacements in *Drosophila* are driven by positive selection. *J Mol Evol* **57** (Suppl 1): S154–S164.
- Scheffler K, Martin DP, Seoighe C (2006). Robust inference of positive selection from recombining coding sequences. *Bioinformatics* **22**: 2493–2499.
- Scheffler K, Seoighe C (2005). A Bayesian model comparison approach to inferring positive selection. *Mol Biol Evol* **22**: 2531–2540.
- Siltberg J, Liberles DA (2002). A simple covarion-based approach to analyse nucleotide substitution rates. *J Evol Biol* **15**: 588–594.
- Smith JM, Smith NH (1996). Synonymous nucleotide divergence: what is 'saturation'? *Genetics* **142**: 1033–1036.
- Smith NG, Eyre-Walker A (2002). Adaptive protein evolution in *Drosophila*. *Nature* **415**: 1022–1024.
- Soyer OS, Goldstein RA (2004). Predicting functional sites in proteins: site-specific evolutionary models and their application to neurotransmitter transporters. *J Mol Biol* **339**: 227–242.
- Stern A, Pupko T (2006). An evolutionary space-time model with varying among-site dependencies. *Mol Biol Evol* **23**: 392–400.
- Suzuki Y (2004). New methods for detecting positive selection at single amino acid sites. *J Mol Evol* **59**: 11–19.
- Suzuki Y, Gojoberi T (1999). A method for detecting positive selection at single amino acid sites. *Mol Biol Evol* **16**: 1315–1328.
- Tajima F (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123**: 585–595.
- Tellgren A, Berglund AC, Savolainen P, Janis CM, Liberles DA (2004). Myostatin rapid sequence evolution in ruminants predates domestication. *Mol Phylogenet Evol* **33**: 782–790.
- Templeton AR (1996). Contingency tests of neutrality using intra/interspecific gene trees: the rejection of neutrality for the evolution of the mitochondrial cytochrome oxidase II gene in the hominoid primates. *Genetics* **144**: 1263–1270.
- Tuffley C, Steel M (1998). Modeling the covarion hypothesis of nucleotide substitution. *Math Biosci* **147**: 63–91.
- Uzzel T, Corbin KW (1971). Fitting discrete probability distribution to evolutionary events. *Science* **172**: 1089–1096.
- Wang HC, Spencer M, Susko E, Roger AJ (2007). Testing for covarion-like evolution in protein sequences. *Mol Biol Evol* **24**: 294–305.

- Wayne ML, Simonsen K (1998). Statistical tests of neutrality in the age of weak selection. *Trends Ecol Evol* **13**: 1292–1299.
- Whelan S, Goldman N (2001). A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Mol Biol Evol* **18**: 691–699.
- Williams PD, Pollock DD, Blackburne BP, Goldstein RA (2006). Assessing the accuracy of ancestral protein reconstruction methods. *PLoS Comp Biol* **2**: e69.
- Wilson DJ, McVean G (2006). Estimating diversifying selection and functional constraint in the presence of recombination. *Genetics* **172**: 1411–1425.
- Wong WS, Sainudiin R, Nielsen R (2006). Identification of physicochemical selective pressure on protein encoding nucleotide sequences. *BMC Bioinformatics* **7**: 148.
- Wong WS, Yang Z, Goldman N, Nielsen R (2004). Accuracy and power of statistical methods for detecting adaptive evolution in protein coding sequences and for identifying positively selected sites. *Genetics* **168**: 1041–1051.
- Xia X, Li WH (1998). What amino acid properties affect protein evolution? *J Mol Evol* **47**: 557–564.
- Yamaguchi-Kabata Y, 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.
- Yang Z (1993). Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. *Mol Biol Evol* **10**: 1396–1401.
- Yang Z (1996). Among-site rate variation and its impact on phylogenetic analyses. *Trends Ecol Evol* **11**: 367–372.
- Yang Z (1997). PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput Appl Biosci* **13**: 555–556.
- Yang Z (1998). Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. *Mol Biol Evol* **15**: 568–573.
- Yang Z (2006). *Computational Molecular Evolution*. Oxford University Press: Oxford.
- Yang Z, Bielawski JP (2000). Statistical methods for detecting molecular adaptation. *Trends Ecol Evol* **15**: 496–503.
- Yang Z, Nielsen R (2000). Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Mol Biol Evol* **17**: 32–43.
- Yang Z, Nielsen R (2002). Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. *Mol Biol Evol* **19**: 908–917.
- Yang Z, Nielsen R, Goldman N, Pedersen AM (2000). Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics* **155**: 431–449.
- Yang Z, Swanson WJ (2002). Codon-substitution models to detect adaptive evolution that account for heterogeneous selective pressures among site classes. *Mol Biol Evol* **19**: 49–57.
- Yang Z, Wong WS, Nielsen R (2005). Bayes empirical bayes inference of amino acid sites under positive selection. *Mol Biol Evol* **22**: 1107–1118.
- Zhai W, Slatkin M, Nielsen R (2007). Exploring variation in the  $d_N/d_S$  ratio among sites and lineages using mutational mappings: applications to the influenza virus. *J Mol Evol* (in press).
- Zhang J (2004). Frequent false detection of positive selection by the likelihood method with branch-site models. *Mol Biol Evol* **21**: 1332–1339.
- Zhang J, Kumar S, Nei M (1997). Small-sample tests of episodic adaptive evolution: a case study of primate lysozymes. *Mol Biol Evol* **14**: 1335–1338.
- Zhang J, Nielsen R, Yang Z (2005). Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. *Mol Biol Evol* **22**: 2472–2479.
- Zhu L, Bustamante CD (2005). A composite-likelihood approach for detecting directional selection from DNA sequence data. *Genetics* **170**: 1411–1421.