Journal of Structural and Functional Genomics **3:** 201–212, 2003.

Maximum likelihood methods for detecting adaptive evolution after gene duplication

Joseph P. Bielawski[*] & Ziheng Yang
Department of Biology, University College London, Darwin Building, Gower Street, London WCIE 6BT, United Kingdom
[*]*Author for correspondence (e-mail: j.bielawski@ucl.ac.uk)*

Received 24.01.2002; Accepted in final form 29.08.2002

Key words: codon model, ECP, EDN, gene family, maximum likelihood, positive selection, Troponin C

Abstract

The rapid accumulation of genomic sequences in public databases will finally allow large scale studies of gene family evolution, including evaluation of the role of positive Darwinian selection following a duplication event. This will be possible because recent statistical methods of comparing synonymous and nonsynonymous substitution rates permit reliable detection of positive selection at individual amino acid sites and along evolutionary lineages. Here, we summarize maximum-likelihood based methods, and present a framework for their application to analysis of gene families. Using these methods, we investigated the role of positive Darwinian selection in the ECP-EDN gene family of primates and the Troponin C gene family of vertebrates. We also comment on the limitations of these methods and discuss directions for further improvements.

Introduction

Duplication of genetic material is generally accepted as an important precursor of functional divergence (Ohno, 1970; Ohta, 1988a, 1988b; Hughes, 1999). Indeed, the majority of genes in higher organisms are members of multigene families or superfamilies (Hughes, 1999). There is often an acceleration of the non-synonymous rate following gene duplication (Li, 1985; Lynch and Conery, 2000), but the mechanism for this acceleration is not clear. Studies of several gene families indicated that natural selection accelerated the fixation rate of nonsynonymous substitutions shortly after a duplication event, presumably to adapt those proteins to a new or modified function (Zhang *et al.*, 1998; Schmidt *et al.*, 1999; Duda and Palumbi, 1999; Rooney and Zhang, 1999; Bielawski and Yang, 2000). However, an accelerated nonsynonymous rate also could be driven by a relaxation, but not complete loss, of selective constraints. Here, duplicated proteins evolve under relaxed functional constraints for some period of time, after which functional divergence occurs when formerly neutral substitutions convey a selective advantage in a novel environment or genetic background; this model was named the "Dykhuizen-Hartl effect" by Zhang *et al.* (1998). The rapid accumulation of genomic sequences in public databases will allow large scale studies of functional divergence in gene families.

For protein coding genes, the most compelling evidence for positive Darwinain selection is derived from comparison of nonsynonymous (amino acid replacement) and synoymous (silent) substitution rates, d_N and d_S, respectively. The difference between these two rates, measured as the ratio $\omega = d_N/d_S$, reflects the effect of selection on the protein product of the gene (Kimura, 1983). For example, if nonsynonymous mutations are deleterious, purifying selection will reduce or prevent their fixation rate and d_N/d_S will be less than 1, whereas if nonsynonymous mutations are neutral then they will be fixed at the same rate as synonymous mutations and $d_N/d_S = 1$. Only under positive Darwinian selection can nonsynonymous mutations be fixed at a rate higher than that of synonymous substitutions, with $d_N/d_S > 1$. Traditionally, to demonstrate positive Darwinian selection models of neutral evolution and purifying selection must be rejected, *i.e.*, the d_N/d_S ratio must be shown to be significantly greater than 1 (Hughes and Nei, 1988; Yang, 1998).

Models of adaptive evolution by gene duplication (Ohta, 1988a, 1988b; Hughes, 1999) make clear predictions about patterns of genetic changes. After duplication, natural selection favours the fixation of mutations in one or both copies that adapt them to divergent functions. Once new or enhanced functions become established, positive selection ceases and purifying selection acts to maintain the new functions. For protein coding genes, this means nonsynonymous substitutions will be accelerated following the duplication, and then slow down due to increased effects of purifying selection. Furthermore, many amino acids in a protein are under strong structural constraints, and adaptive evolution following a duplication event will most likely operate at only a subset of sites. An ω ratio measured as an average over all sites and evolutionary time will rarely be greater than one, and is too stringent a criterion for detecting positive selection (*e.g.*, Ward *et al.*, 1997; Crandall *et al.*, 1999; Bielawski and Yang, 2001). Hence, the apparent period of relaxed selection following gene duplication (*e.g.*, Lynch and Conery, 2000) might also reflect adaptive divergence of duplicate genes at just a subset of amino acid sites.

Recent ML models of codon substitution relax the assumption of a single ω ratio for all branches of a phylogeney (Yang, 1998). Theses models can provide a framework for constructing likelihood ratio tests of changes in selective pressure following gene duplication (Bielawski and Yang, 2001). Other codon models allow the ω ratio to vary among amino acid sites (Nielsen and Yang, 1998; Yang *et al.*, 2000). Very recently, Yang and Nielsen (2002) developed a third type of model that can simultaneously account for variation in selective constraints among sites and lineages. Here, we summarize all three types of models, and present a framework for their application to gene families. We also comment on limitations of current methods and discuss directions for further improvements.

ML estimation of the d_N/d_S ratio

Markov model of codon evolution

A markov process is used to describe substitutions between 61 of the 64 sense codons. The three stop codons are excluded because they aren't allowed within a protein. Independence among the codon sites of a gene is assumed, and hence the substitution process can be considered one codon site at a time. For any single codon site, the model describes the instantaneous substitution rate from codon i to codon j, q_{ij}.

In the following, we describe the basic model of Goldman and Yang (1994). A similar, but simpler model was described by Muse and Gaut (1994). Because transitional substitutions are known to occur more often than transversions, the rate is multiplied by the κ parameter when the change involves a transition; this is the transition/transversion rate ratio. Usage of codons within genes also can be highly biased, and consequently, the rate of change from i to j is multiplied by the equilibrium frequency of codon j (π_j). Finally, selective constraints acting on substitutions at the amino acid level affect the rate of change when that change represents a nonsynonymous substitution. The rate is multiplied by the ω parameter if the change is nonsynonymous; the ω parameter is the nonsynonymous/synonymous rate ratio (d_N/d_S).

The substitution model is specified by the instantaneous rate matrix, $Q = \{q_{ij}\}$, where

$$q_{ij} = \begin{cases} 0, & \text{if } i \text{ and } j \text{ differ at two or three codon positions} \\ \mu\pi_j & \text{if } i \text{ and } j \text{ differ by a synonymous transversion} \\ \mu\kappa\pi_j, & \text{if } i \text{ and } j \text{ differ by a synonymous transition} \\ \mu\omega\pi_j, & \text{if } i \text{ and } j \textit{ differ} \text{ by a nonsynonymous transversion} \\ \mu\omega\kappa\pi_j, & \text{if } i \text{ and } j \text{ differ by a nonsynonymous transition.} \end{cases}$$

The diagonal elements of the matrix Q are defined by the mathematical requirement that the row sums are equal to zero. Because separate estimation of the rate (μ) and time (t) is not possible, the rate (μ) is fixed so that the expected number of nucleotide substitutions per codon is equal to one. This scaling allows us to measure time (t) by the expected number of substitutions per codon, *i.e.*, genetic distance. The probability that codon i is substituted by codon j after time t is $p_{ij}(t)$, and $\mathrm{P}(t) = \{p_{ij}(t)\} = e^{Qt}$.

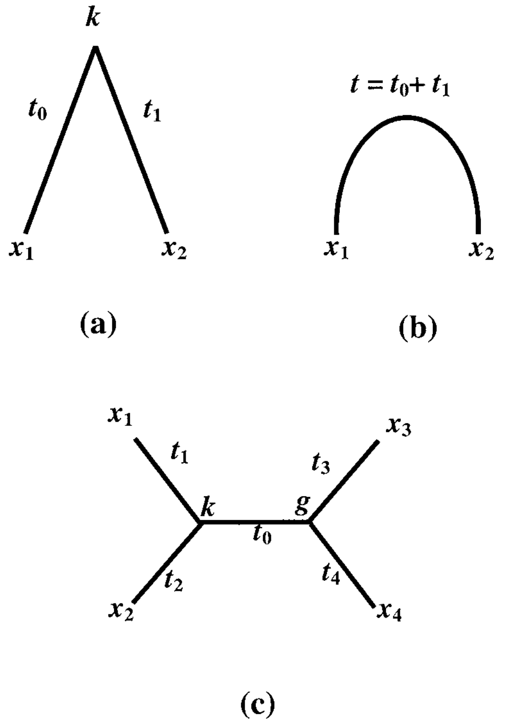


Figure 1. Trees (a) and (b) are for two sequences; (a) is rooted and (b) is unrooted. Codon models are time-reversible; hence, the root cannot be identified and only the sum of branch lengths, t, is estimable. Tree (c) is an unrooted tree for four sequences.

ML estimation on a phylogeny

Let us consider a simple case, the likelihood of observing the data of a pair of species. Suppose there are n codon sites in a gene, and a certain site (h) has codons CCC and CTC. The data at site h, denoted $\mathbf{x}_h = \{CCC, CTC\}$ are related to an ancestor with codon k by branch lengths t_0 and t_1 (Fig. 1a). The probability of site h is

$$L(\mathbf{x}_h) = \sum_k \pi_k p_{k,CCC}(t_0) p_{k,\mathrm{CTC}}(t_1).$$

Since the ancestral codon is unknown the summation is over all 61 possible codons for k.

The log likelihood is a sum over all codon sites (n) in the sequence

$$l(t,\kappa,\omega) = \sum_{h=1}^{n} \log\{\mathrm{L}(\mathbf{x}_h)\}.$$

Codon frequencies (π_i's) are usually estimated using observed base or codon frequencies. Parameters ω, κ, and t are estimated by maximizing the likelihood function. Because the root cannot be identified, t_0 and t_1 cannot be estimated individually, and only $t_1 + t_2 = t$ is estimated (Fig. 1b). Since an analytical solution is not possible, numerical optimisation algorithms are used.

Likelihood calculation for multiple lineages on a phylogeny (*e.g.*, Fig. 1c) is an extension of the calculation for two lineages. As in the case of two sequences, the root cannot be identified and is fixed at one of the ancestral nodes arbitrarily. For example, given an unrooted tree with four species and two ancestral codons, k and g, the probability of observing the data at codon site h, $\mathbf{x}_h = \{x_1, x_2, x_3, x_4\}$ (Fig. 1c), is

$$\mathrm{L}(\mathbf{x}_h) = \sum_k\sum_g [\pi_k p_{kx_1}(t_1) p_{kx_2}(t_2) p_{kg}(t_0) p_{gx_3}(t_3) \\ p_{gx_4}(t_4)]$$

The quantity in the brackets is the contribution to the probability of observing the data by ancestral codons k and g at the two ancestral nodes. For an unrooted tree of N species, with N–2 ancestral nodes, the data at each site will be a sum over $61^{(N-2)}$ possible combinations of ancestral codons. The log likelihood is a sum over all codon sites in the alignment

$$l = \sum_{h=1}^{n} \log\{L(\mathbf{x}_h)\}.$$

As in the two-species case, numerical optimisation is used to estimate ω, κ, and the (2N–3) branch length parameters (t's).

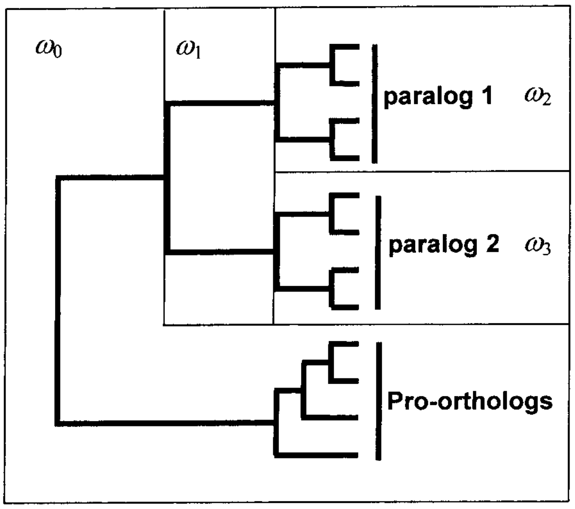
Detecting lineage-specific changes in selective pressure

Models of functional divergence in gene families emphasize the episodic nature of substitution rates (Ohta, 1988a, 1988b; Hughes, 1999). Under these models, most evolution in gene families will be by purifying selection, with episodes of either reduced selective constraints or positive Darwinian selection following duplication events. If divergence is driven by positive Darwinian selection, nonsynonymous mutations might be fixed at a much higher rate than synonymous mutations immediately following the duplication event. However, if the gene family evolves under purifying selection at other times, comparisons among distantly related pairs of sequences are unlikely to yield d_N/d_S ratios greater than 1. For this reason, adaptive evolution will be very difficult to detect in gene families by using a pairwise approach to estimating d_N/d_S ratios.

We need tools to (1) estimate selective pressures at different time points in the phylogenetic history of a gene, and (2) test the hypothesis that those estimates differ significantly. Codon models that allow independent ω ratios in different parts of a phylogeny provide a framework for analysing changes in selective pressure over time (Yang, 1998; Bielawski and Yang, 2001). The advantage of this likelihood-based approach is that it does not depend on the accuracy of hypothetical ancestral sequences, as in other approaches, although it does incur greater computational costs (Yang, 2001).

Models of variable selective pressures among branches

The null model (the one-ratio model) assumes the same ω ratio for all branches (Fig. 2; R1). Nested models are constructed based on the assumption that selective constraints change following a gene duplication (Fig. 2). Here, the likelihood calculation is modified so that independent ω's are used to calculate rate matrices (Q) and transitions probabilities for different branches. In a simple case where there is only a single duplication event in a phylogeny, model R2 (two-ratio model) assumes two independent ω ratios: one ratio for all branches predating a duplication event and a second for all branches postdating a duplication event (Fig. 2; R2). A likelihood ratio test (LRT) of the one-ratio model with model R2 examines the difference between average selective constrains before and after a duplication event. A more complex model assumes three independent ω ratios: one for all branches predating the duplication event, a second for the branches immediately following the duplication event, and a third for all subsequent branches (Fig. 2; R3). An LRT comparing model R2 and R3 examines the difference between selective constraints at branches immediately after the duplication and those in subsequent branches. Finally, R4 extends R3 to allow selective pressure to differ between paralogous genes (Fig. 2; R4). Different hypotheses, perhaps involving multiple duplication events, also can be constructed; we used this example (Fig. 2) to illustrate the general framework of the ML approach.



One ratio (R1): $\omega_0 = \omega_1 = \omega_2 = \omega_3$
Two ratios (R2): $\omega_0 \neq \omega_1 = \omega_2 = \omega_3$
Three ratios (R3): $\omega_0 \neq \omega_1 \neq \omega_2 = \omega_3$
Four ratios (R4): $\omega_0 \neq \omega_1 \neq \omega_2 \neq \omega_3$

Figure 2. Phylogeny for a hypothetical gene family. Pro-orthologs are the single-copy genes that predate the duplication event. Paralogs 1 and 2 refer to the two sets of genes that derived from the duplication event. The one-ratio model (R1) assumes all branches have the same ω parameter. The R2 model assumes one ω is for all branches that predate the duplication event, and a second ω for all branches that follow the duplication event. Model R3 assumes one ω for all branches that predate the duplication event, a second ω for the branches that immediately follow the duplication event, and a third ω is for subsequent branches. Model R4 is an extension of R3, and allows selective constraints to differ in the two paralogs.

Changes in nonsysnonymous substitution rates following gene duplication in the ECP-EDN gene family

The ECP-EDN gene family of primates is comprised of the eosiophil-derived neurotoxin (EDN) and the eosinophil cationic (ECP) genes. ECP and EDN are ribonucleases present in the large specific granules of eosinophilic leukocytes (Rosenberg and Domachowske, 1999). While both ECP and EDN have host-defence roles, their specific functions differ. ECP is a cationic toxin, apparently functioning as a non-specific toxin to parasites and bacteria (Rosenberg and Domachowske, 1999). EDN is considerably less cationic, but has strong ribonucleic activity making it a potent antiviral agent via ribonucleic degradation of viral RNA (Rosenberg and Domachowske, 1999). Zhang *et al.* (1998) studied the evolution of this gene

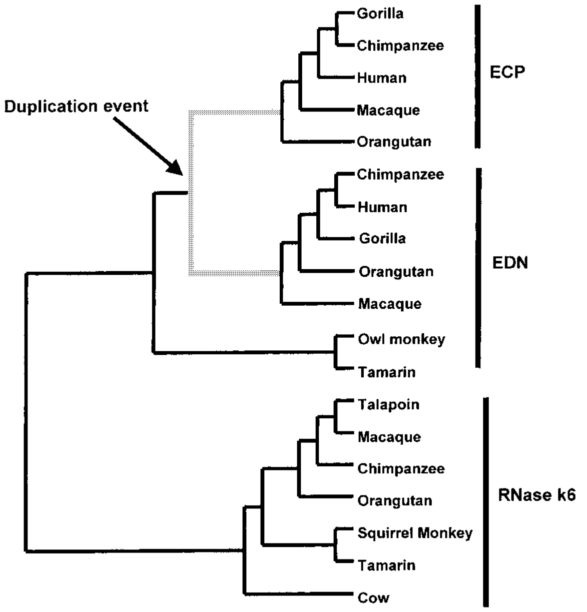


Figure 3. A phylogenetic tree for the ECP-EDN gene family. The tree is rooted with RNase k6 genes. All analyses were conducted using unrooted topologies; this topology is rooted for convenience. Note that branch lengths are not to scale.

family and found an excess of nonsynonymous substitutions over synonymous substitutions in the branch leading to the ECP gene. They suggested that the anti-parasitic function of ECP evolved shortly after the duplication that gave rise to ECP and EDN. We use the ECP-EDN family to demonstrate the application of codon models to studies of evolution by gene duplication. A phylogenetic hypothesis for the ECP-EDN gene family is presented in Figure 3.

We estimated ω as an average over all sites and branches (Fig. 3) and the ratio was substantially less than 1 (one ratio model; $\omega = 0.51$). The one-ratio model was compared with model R2, and the LRT indicated that model R2 provided a significantly better fit to these data ($2\delta = 14.92$, d.f. = 1, $P = 0.0001$). Estimate of ω under model R2 indicated a significant increase in the average rate of nonsynonymous substitution following the duplication event (Table 1; $\omega_0 = 0.38$ vs. $\omega_1 = 0.87$). To test if more recent selective pressure in ECP and EDN differed from that immediately following the duplication event, we compared model R2 with R3. The likelihood of model R3 was significantly better than R2 ($2\delta = 4.2$, d.f. = 1, $P = 0.04$), with estimates of ω's indicating that adaptive evolution had occurred subsequent to the duplication event (Table 1; $\omega_1 = 2.00$). Lastly, we investigated the assumption of the same selective pressures for both EDN and ECP by comparing model R3 with R4 . Likelihood of model R4 was significantly better than R3 ($2\delta = 9.64$, d.f. = 1, $P = 0.002$), indicating different selective pressures in ECP and EDN. Interestingly, parameter estimates indicated positive Darwinian selection for the ECP clade ($\omega_3 = 1.56$) and purifying selection in the EDN clade ($\omega_2 = 0.45$).

Table 1. Parameter estimates and likelihood scores for the ECP-EDN gene family under different branch-specific models

Model	NP	Parameters for branches [a]	*l*
One-ratio (R1)	1	$\omega_0 = 0.51$ $\omega_1 = \omega_0$ $\omega_2 = \omega_0$ $\omega_3 = \omega_0$	−3109.63
Two-ratios (R2)	2	$\omega_0 = 0.38$ $\omega_1 = 0.87$ $\omega_2 = \omega_1$ $\omega_3 = \omega_1$	−3102.17
Three-ratios (R3)	3	$\omega_0 = 0.38$ $\boldsymbol{\omega_1} = \mathbf{1.99}$ $\omega_2 = 0.72$ $\omega_3 = \omega_2$	−3100.06
Four-ratios (R4)	4	$\omega_0 = 0.38$ $\boldsymbol{\omega_1} = \mathbf{2.00}$ $\omega_2 = 0.45$ $\boldsymbol{\omega_3} = \mathbf{1.56}$	−3095.50

[a] Models R1 to R4 are presented in Figure 2. The topology for the ECP-EDN gene family is presented in figure 3: ω_0 is for branches that predate the duplication event, $\omega\ _1$ for the branches that immediately follow the duplication event (grey branch in Fig. 3), ω_2 is for branches in the EDN clade excluding Owl monkey and Tamarin, and ω_3 is for all branches in the ECP clade. NP is the number of freely estimated ω ratios.

Our findings indicate there was a significant increase in the rate of fixation of nonsynonymous substitutions following the duplication that gave rise to the ECP-EDN gene family. This rate increase was partially due to adaptive evolution immediately following the duplication. Presumably, natural selection increased the fixation rate for nonsynonymous mutations that adapted ECP for greater anti-parasitic activity, and nonsynonymous mutations that enhanced anti-viral ribonuclease activity in EDN. Our findings also suggest that ECP has continued to evolve under positive Darwinian selection long after the initial period of function divergence. Rosenberg and Domachowske (1999) have speculated that both EDN and ECP might have acquired specialized anti-viral activity, perhaps against respiratory viral pathogens. Perhaps subsequent adaptive evolution in the ECP clade reflects long term selective pressure for effective antiviral activity against respiratory viral pathogens.

Identification of amino acid sites under adaptive evolution

In general, most amino acid sites are subject to strong functional constraints, with d_N, and consequently ω, close to zero (Sharp, 1997). Most studies of molecular evolution indicate adaptive changes occur at only a subset of sites (*e.g.*, Golding and Dean, 1998); hence, use of ω averaged over all sites has little power to detect positive selection (*e.g.*, Endo *et al.*, 1996). In the previous section we assumed that all amino acid sites were subject to identical selective pressure, with a single ω ratio applied to all sites. By allowing an independent ω for specific intervals of time, such as for those branches that immediately postdate a duplication event, we greatly increased the power to detect positive selection. However, finding adaptive evolution during a certain time interval provided no information about variation in selective constraints among sites. Hence, we also need tools (1) to test for evolution by positive Darwinian selection at a subset of codons, and (2) to identify such sites when they exist.

Two strategies can be taken. In the first, amino acid sites are classified into several independent ω ratio classes based on prior knowledge of structural and functional domains. Likelihood calculation in this case is similar to that under the model of one ω ratio for all sites, except that different ω parameters are used to calculate the transition probabilities at different sites (Yang, 2001; Yang and Swanson 2002). Under the second strategy, information of structural and functional domains of the protein is unknown, or not used, and a statistical distribution is used to account for variation of the ω ratio among sites (Nielsen and Yang, 1998; Yang *et al*., 2000). After ML estimates of parameters in the ω distribution are obtained, an empirical Bayes approach is used to predict the most likely ω class for each site, with the posterior probability providing a measure of reliability (Nielsen and Yang, 1998; Yang *et al*., 2000). It is important to point out that these codon models assume a constant selection pressure along the tree and are thus conservative in detecting positive selection. Despite this assumption, they proved useful for identifying sites under positive selection and studying the process of adaptive molecular evolution (*e.g*., Zanotto *et al.,* 1999; Bishop *et. al*., 2000; Haydon *et al*., 2001; Swanson *et al*., 2001).

Models of variable selective pressure among sites

Usually, the structural and functional domains of a protein are not well understood or are completely unknown. For this reason we focus on models that use a statistical distribution to account for variation of the ω ratio among sites. Collectively, Nielsen and Yang (1998) and Yang *et al.* (2000) implemented 12 such models; here we discuss five of them (M1, M2, M3, M7, and M8) because they comprise the set generally recommended for data analysis (Yang *et al*., 2000; Anisimova *et al*., 2001).

The first two models, introduced by Nielsen and Yang (1998), specify just a few discrete ω classes. The 'neutral' model (M1) assumes two classes of sites with fixed ω values; in one class of sites nonsynonymous mutations are completely selected against, with $\omega_0 = 0$, and the other class is comprised of neutral sites, with $\omega_1 = 1$. The proportion of sites is estimated via ML for only one ω class (p_0), as $p_1 = 1 - p_0$. The 'selection' model (M2) adds a third class of sites with the underlying ω ratio freely estimated from the data. M2 has parameters p_0, p_1, and ω_2. These models appear too simple to capture the complexity of the substitution process of various proteins (Yang, 2001). They are conservative for the purpose of testing and identifying sites under positive selection (Anisimova *et al.*, 2001).

Yang *et al.* (2000) introduced more general models M3, M7, and M8. M3 (discrete) assumes k site classes, with corresponding ω_i and p_i for each class ($i \ = 1, \dots k$) estimated as parameters. M7 (beta) assumes that ω ratios are distributed among sites according to a beta distribution. Note that the beta distribution is very flexible, taking a variety of shapes within the interval (0,1) depending on the parameters p and q. M8 (beta&ω) is an extension of M7, having an extra class of sites with an independent ω ratio freely estimated from the data.

An LRT of the one-ratio model with M3 is a test of variable selective pressures among sites. To spe-

Table 2. Parameter estimates and likelihood scores for the ECP-EDN gene family under different models of variable ω ratios among sites.

Model	Parameter estimates	Positively selected sites	l
One-ratio	$\omega = 0.79$	None	−1912.05
Neutral (M1)	$(\omega_0 = 0), f_0 = 0.36$ $(\omega_1 = 1), (f_1 = 0.64)$	Not allowed	−1890.38
Selection (M2)	$(\omega_0 = 0), f_0 = 0.36$ $(\omega_1 = 1), f_1 = 0.38$ $\boldsymbol{\omega_2 = 2.22}$, $(\boldsymbol{f_2 = 0.26})$	37 sites[a]	−1886.59
Discrete (M3)	$\omega_0 = 0.17, f_0 = 0.56$ $\boldsymbol{\omega_1 = 1.9}$ **,** $(\boldsymbol{f_1 = 0.44})$	70 sites[b]	−1885.49
Beta (M7)	$p\ = 0.011,\ q\ = 0.05$	Not allowed	−1891.10
Beta&ω (M8)	$p = 20,\ q = 99$ $f_0 = 0.56$ $\boldsymbol{\omega_1 = 1.9}$, $(\boldsymbol{f_1 = 0.44})$	70 sites[b]	−1885.50

[a] None of the positive selection sites identified under model M2 had posterior probabilities >0.95.

[b] Positive selection sites identified under M3 and M8 were identical; sites with posterior probabilities ≥ 0.95 under M8 were as follows (sites with posterior probabilities $\geq$ 0.99 are in bold): **19M**, 27A, **28R**, **30P**, **34R**, **39A**, **44S**, **45L** , **48P**, **49R**, **52I**, 55R, **62W**, 72R, **77N**, **85Q**, 86S, **88R**, **93R**, 94T, 96N, **100R**, 102R, **103F**, 117A, 118Q, **124T**, 126A, **127D**, **129P**, 130G, 143P, 144?, 159T. Letters refer to the amino acid residue found in the human ECP gene.

cifically test for a portion of sites evolving by positive Darwinian selection, LRTs are conducted to compare M1 with M2 and M7 with M8. Positive selection is indicated when a freely estimated ω parameter is >1 and the LRT is significant.

Detection of positive selection sites in the EDN-ECP gene family

We now re-examine the EDN-ECP dataset using models of variable ω ratios among sites. Because previous analysis indicated positive selection followed the duplication event, and because the RNase-k6 sequences were highly divergent, we excluded the outgroups from these analyses, leaving twelve sequences in the dataset. Note that a very high divergence can reduce power to detect positive selection under models of variable ω ratios among sites (Anisimova *et al.*, 2001). Results obtained under several models are presented in Table 2.

Averaging ω over all sites and branches for these sequences gives an ω ratio of 0.818, an average which indicates evolution by weak purifying selection. However, a LRT of the one-ratio model with M3 ($k = 3$) indicates that selective pressure is not uniform among sites ($2\delta = 53.12$, d.f. = 4, $P < 0.0001$). Furthermore, all three models which permit an ω parameter to exceed 1 (M2, M3, and M8) indicated a large fraction of sites evolving under positive Darwinian selection (Table 2). LRTs indicate that models M2 and M8 fit the data significantly better than models M1 and M7, in which positively selected sites are not allowed (M1 v. M2: $2\delta = 7.58$, d.f. = 2, $P = 0.02$; M7 v. M8: $2\delta = 11.2$, d.f. = 2, $P = 0.004$). Clearly, some variation in selective pressure is due to evolution by positive selection. It is worth noting that parameter estimates also suggest a large fraction of sites evolving under strong purifying selection (Table 2). Clearly, estimation of the single ω parameter, as an average over sites, did not provide a sensible measure of selective constraints for the ECP-EDN gene family.

Zhang *et al.* (1998) reported that the number of arginines increased substantially along the branch leading to the ECP clade. We wanted to evaluate the pattern of arginine evolution in the ECP clade at sites subject to positive selection. The Bayes theorem was used to identify candidate positive selection sites. There was no major increase in the number of argenines at those sites within the ECP clade. This suggests that more recent adaptive pressure in ECP might differ from the early selection pressure to enhance cationic toxicity. Interestingly, residues critical to ribonucleic activity in ECP are evolutionary conserved despite their making no contribution to the cationic toxicity of ECP (Rosenberg and Domachowske, 1999).

It is important to note that the current Baysean analysis did not distinguish between sites related to adaptive divergence of ECP and EDN, and those sites that have been the target of continual adaptive pressure in ECP. Additional sequences of ECP will be needed to infer positive selection sites specific to that gene. Bayes identification of positively selected sites in EDN and ECP, combined with reconstruction of ancestral amino acid sequences and site-directed mutagenesis, could help identify the key sites associated with EDN and ECP activities.

Lineage-specific changes in selective pressure at specific amino acid sites

Yang and Nielsen (2002) recently extended previous codon models to permit the ω ratio to vary both among lineages and among sites. These new models are called "branch-site" models. Because they were developed to identify adaptive evolution in pre-specified branches, they are suited to studying the role of positive selection in gene families.

We are interested in knowing if adaptive evolution has occurred in the lineages immediately following a duplication event. The branch, or branches, of interest are called the "foreground" branches, and all others are called "background" branches. In the branch-site models, selective constraints are assumed to vary among sites, but at a subset of sites selective constraints are permitted to change in the foreground lineage. Sites changing in the foreground lineage are permitted to have $\omega > 1$. The basic model has four ω site classes. The first two classes of sites, with ω_0 and ω_1, are uniform over the entire phylogeny. The other two classes allow some sites with ω_0 and ω_1 to change to positive selection (ω_2) in the foreground lineage, *i.e.*, sites where $\omega_0 \rightarrow \omega_2$ and $\omega_1 \rightarrow \omega_2$.

Yang and Nielsen (2002) implemented two versions of branch-site models (A and B). In model A, ω_0 is fixed to 0 and ω_1 is fixed to 1; hence positive selection is permitted only in the foreground branch. This model is compared with model M1 (neutral) with d.f. = 2. In model B, ω_0 and ω_1 are free parameters. Thus some sites could evolve by positive selection across the entire phylogeny, whereas other sites could evolve by positive selection in just the foreground branch. Model B is compared with M3 (discrete) with $k = 2$ site classes and d.f. = 2.

Application to the Troponin C gene family

We use the Troponin C gene family to demonstrate the application of the branch-site models (Yang and Nielsen, 2002) to studies of evolution by gene duplication. Through association with actin and tropomyosin, Troponin C inhibits actomysin interaction at submicromolar Ca^{2+} concentrations, and stimulates the interaction at micromolar concentrations (Farah and Reinach, 1995). In vertebrates there was a single duplication in this family that initiated evolution of the two distinct muscle isoforms for Troponin C: (i) the fast skeletal-muscle isoform, and (ii) the cardiac and slow skeletal-muscle isoform. Ohta (1994) examined estimates of d_N and d_S within and between the different isoforms of Torponin C and found a higher rates of nonsynonymous substitution between isoforms than within an isoform. Ohta (1994) considered this as evidence for an accelerated nonsynonymous rate following the duplication event and suggested divergence by positive selection. However, in no comparisons were d_N/d_S ratios greater than 1, and the results also seemed compatible with reduced purifying selection after gene duplication. Here we investigate the evolutionary forces associated with the functional divergence of these two isoforms using branch models, sites models, and branch-site models. The phylogeny for the Troponin C gene family is presented in Figure 4. We treat the branch labelled A as the foreground branch and all other branches in the phylogeny as background branches.

Averaging ω over all sites and branches yielded the estimate $\omega = 0.017$ (one ratio model; Table 3), indicating the overwhelming role of purifying selection in this gene family. We relax the assumption of a single ω for all branches in the two-ratios model. Here we assign one ω ratio for the branch separating the two isoforms (ω_0) and a second for all other branches (ω_1). The branch separating the two isoforms (Fig. 4, branch A) immediately postdates the Troponon C duplication event. Parameter estimates suggested a higher relative rate of nonsynonymous substitution in branch A (Table 3), but the two-ratio model did not provide a significantly better fit to these data (Table 4).

We also applied sites models (Nielsen and Yang, 1998; Yang *et al.*, 2000) to Troponoin C (Table 3).We employed the following model pairs: (i) one-ratio (M0) and M3; (ii) M1 and M2; and (iii) M7 and M8. M3 fitted these data significantly better than the one-ratio model (Table 4), but no parameter estimate

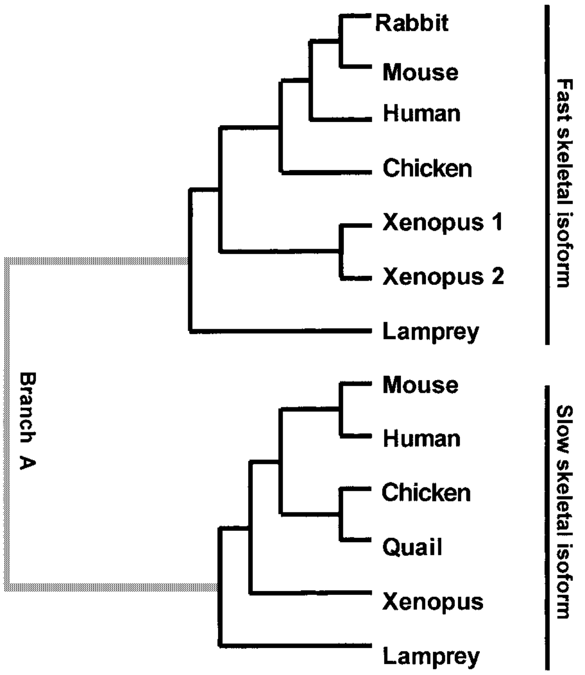


Figure 4. A phylogenetic tree for the Troponin C gene family. Note branch lengths are not to scale. Branch A, indicated in grey, represent the lineage that immediately follows the duplication event in Troponin C. All analyses were conducted using unrooted topologies; this topology is rooted for convenience.

indicated positive selection (Table 3). This finding indicates significant variation in selective constraints among sites, but not by positive Darwinian selection. Comparisons of M1 with M2 and M7 with M8 are consistent with this interpretation, in that both models that permit an ω ratio to exceed 1 (M2 and M8) also failed to indicate sites under positive selection (Table 3).

Branch-site models were applied to Troponin C with branch A (Fig. 4) specified as the foreground branch. In contrast, the new branch-site models indicated evolution by positive Darwinian selection in the Troponin C gene family. Both models A and B indicated a relatively large fraction of sites (model A: 14%; model B: 22%) evolving under positive selection in the foreground branch. Comparison of model A with its null model (M1), and model B with its null model (M3; $k = 2$) indicated that the branch-site models provided a significantly better fit to these data (Table 4). Positively selected sites identified by the Bayes method were listed in Table 3. Model B identified fewer sites than model B (Table 3).

Table 4. Likelihood ratio test statistics (2δ) for the Troponin C dataset.

	2δ	df	*P*-value [a]
LRT of ω at branch A(Fig. 4)			
One ratio vs. two ratios	1.02	1	0.31
LRTs of variable ω's among sites			
One ratio vs. M3	88.62	2	<0.0001
M1 vs. M2	679.14	2	<0.0001
M7 vs. M8	2.70	2	0.26
LRTs of variable ω's along branch A (Fig. 4)			
M1 vs. Model A	152.54	2	<0.0001
M3 vs. Model B	13.86	2	0.001

Initial application of the branch-sites models to two single copy genes (primate lysozme and tumour suppressor BRCA1) yielded mixed results, and Yang and Nielsen (2002) speculated that branch-site models might not provide a significant improvement over previous models in many single copy genes. In contrast, our analysis of the Troponin C gene family illustrates the utility of these new models. Previous codon models did not have the power to detect adaptive evolution in Troponin C because adaptation appears to have occurred at a subset of sites for a limited period of time. This mode of molecular adaptation is likely to be very common in gene families. Indeed, these new models also were successful in detecting adaptive evolution after the gene duplication of the cone opsin that gave rise to rhodopsin genes of vertebrates, a gene family in which the sites-models and branch-models were not effective (B.S.W. Chang, personal communication).

Power and accuracy of ML methods

Initial application of these codon based models led to the detection of positive selection in many genes for which it had not previously been known. In particular, relaxing the assumption of a constant selective pressure over sites appears to increase the power of the detection method considerably. For example, Yang *et al.* (2000) analysed 10 genes from a variety of genomes and identified six of them to be under positive selection. Moreover, previous studies using average d_N and d_S over all sites and branches indicated that evolution in the *nef* gene of HIV1 (Pilkat *et al.*, 1997) and the *DAZ* gene in primates (Agulnick *et al.*, 1998) was dominated by genetic drift, yet sub-

Table 3. Parameter estimates and likelihood scores for the Troponin C gene family under different branch models, sites models, and branch-site models.

Model	p	Estimates of ω parameters	Positive selection	l
One-ratio	1	$\omega = 0.02$	None	−2945.08
Branch specific				
Forground/background	2	$\omega_0 = 0.15$ (branch A; Fig 4) $\omega_1 = 0.02$ (background)	None	−2944.57
Site specific				
Neutral (M1)	1	$(\omega_0 = 0), f_0 = 0.45$ $(\omega_1 = 1), f_1 = 0.55$	Not allowed	−3245.91
Selection (M2)	3	$(\omega_0 = 0), f_0 = 0.38$ $(\omega_1 = 1), f_1 = 0.02$ $\omega_2 = 0.03, (f_2 = 0.60)$	None	−2906.34
Discrete (M3), $k = 2$		$\omega_0 = 0.01, f_0 = 0.76$ $\omega_1 = 0.08, (f_1 = 0.24)$	None	−2900.77
Beta (M7)	2	$p = 0.42, q = 17.08$	Not allowed	−2898.82
Beta&ω (M8)	4	$p = 0.25, q = 8.22$ $f_0 = 0.69$ $\omega \ = 0.015, (f_1 = 0.31)$	None	−2897.47
Branch-site				
Model A	3	$(\omega_0 = 0), f_0 = 0.43$ $(\omega_1 = 1), f_1 = 0.33$ $\boldsymbol{\omega_2 = 89, (f_{2+3} = 0.14)}$	**Foreground:** **22 sites**[a] Background: Not allowed	−3169.63
Model B	5	$\omega_0 = 0.01, f_0 = 0.56$ $\omega_1 = 0.07, f_1 = 0.22$ $\boldsymbol{\omega_2 = 76, (f_{2+3} = 0.22)}$	**Foreground:** **12 sites**[b] Background: None	−2893.84

[a] Positive selection sites identified under Model A with posterior probabilities ≥ 0.95 (sites with posterior probabilities ≥ 0.99 are in bold): **15V,** 16E, 17Q, **23K**, **32I**, **34V,** 35L, 37A, 39D, **41C, 64Q, 65E, 66M, 68D, 78V, 90C, 95S, 109M, 126M,** 157Y, **163F**, **165K**. Letters refer to the amino acid residue found in the slow skeletal-muscle isoform of Human.

[b] Positive selection sites identified under Model B with posterior probabilities ≥ 0.95 (sites with posterior probabilities ≥ 0.99 are in bold): 11Y, **15V**, 17Q, 23K, 24N, 35L, **41C**, **60P**, **64Q**, **90C,** 95S, **163F**. Letters refer to the amino acid residue found in the slow skeletal-muscle isoform of Human.

sequent ML analyses indicated a small fraction of sites evolving under positive selection and the remaining sites under purifying selection (*nef*, Nielsen and Yang, 1998; Zanotto *et al*., 1999; *DAZ*, Bielawski and Yang, 2000).

Although successful in those cases, the accuracy and power of the LRT and Bayesian site identification were largely unknown. Anisimova *et al.* (2001) used computer simulation to address the performance of the site models (Nielsen and Yang, 1998; Yang *et al*., 2000), leading to the following observations: (i) the LRT statistic (2δ) does not follow the χ^2 distribution because some model parameters are fixed at the boundary of parameter space, and use of the χ^2 distribution made the test conservative; *i.e.*, false positives always occurred less frequently than expected under the specified significance level of the test. (ii) Despite being conservative, the LRT still provided a powerful means of detecting positive selection given enough variation in the data; power was 100% with datasets of 17 taxa, but quite low for datasets of only 5 or 6 taxa. (iii) Sequence length, sequence divergence, and strength of positive selection all influence power of the LRT, but the number of sampled taxa is the most important. (iv) Power is low for both highly similar and highly divergent sequences, and is highest at intermediate or moderately high divergence. (v) The LRT is robust to the assumed distribution for ω across sites.

Anisimova *et al.* (2002) also used computer simulation to study the power and accuracy of Bayesian identification of sites evolving under positive selection. The findings can be summarized as follows: (i) identification of sites is not feasible from a few highly similar sequences; (ii) accuracy of site identification is most dependent on the numbers of lineages. The number of lineages and the sequence divergence are the most important factors affecting power. (iii) Unlike results for LRT, the assumed distribution for ω across sites can affect both accuracy and power of site identification. Anisimova *et al.* (2002) recommended using multiple models to identify positively selected sites.

Limitations and Future prospects

The site models are computationally complex and intensive. Models M2 and M8 are known to have multiple local optima in some datasets. Because accuracy is sensitive to parameter estimation, suboptimal parameter estimates, based on a local optimum, could increase the probability of a type-I error. To avoid being trapped at a local optimum, it is important to run M8 multiple times; minimally it should be run with one initial $\omega > 1$ and one initial $\omega < 1$. Results corresponding to the highest likelihood value should be used.

Codon models are limited in effectiveness by the sample of sequences. Typically, the power to detect adaptive evolution will be low for sequences less than one hundred codons in length. Furthermore, the analysis is impacted by the requirement that information is drawn from both synonymous and nonsynonymous substitutions. Hence, the power can be quite low for highly similar and highly divergent sequences (Anisimova *et al.*, 2001). Although sampling additional lineages can improve power when sequence divergence is low or high, the window of suitable sequence divergence is expected to be narrower than for phylogeny reconstruction.

Gu (1999, 2001) developed a maximum likelihood approach to testing for functional divergence that is based on amino acid substitution rates. The method can detect specific amino acid residues that contribute to functional divergence following gene duplication (Gu, 1999, 2001). Although amino acid substitution rate might not be as sensitive a measure of selective pressure as the d_N/d_S ratio, amino acid evolution is much slower than nucleotide evolution, and Gu's (1999, 2001) approach should provide an excellent tool for studying evolution by gene duplication in gene families too divergent for analysis with codon models.

An interesting avenue for future development of codon models is the “covarion-like” model (Galtier, 2001). Covarion-like models allow site specific rates to vary among lineages, and can be achieved by adding as few as two additional parameters to non-covarion models (Galtier, 2001). Galtier's (2001) ML implementation for nucleotide models is promising, as it revealed a significant amount of site-specific rate variation in ribosomal RNA sequences. Covarion-like codon models might provide an alternative framework for investigating episodic adaptive evolution acting on a few amino acid sites (Galtier, 2001). This approach would have the advantage of not requiring *a priori* knowledge of the specific branches where episodes of adaptive evolution have occurred.

References

Agulnik, A.I., Zharkikh, A., Boettger-Tong, H., Bourgeron, T., McElreavey, K. and Bishop, C.E. (1998) Evolution of the DAZ gene family suggests that Y-linked *DAZ* plays little, or a limited, role in spermatogenesis but underlines a recent African origin for human populations. *Hum. Mol. Genet.* **7,** 1371–1377.

Anisimova, M., Bielawski, J.P. and 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, J.P. and Yang, Z. (2002) Accuracy and power of Bayesian prediction of amino acid sites under positive selection. *Mol. Biol. Evol.* **19**, 950–958.

Bielawski, J.P. and Yang, Z. (2000) Positive and negative selection in the *DAZ* gene family. *Mol. Biol. Evol.* **18,** 523–528.

Bishop,J.G., Dean, A.M. and Mitchell-Olds, T. (2000) Rapid evolution of plant chitinases: molecular targets of selection in plant-pathogen coevolution. *Proc. Natl. Acad. Sci. USA* **97,** 5322–5327.

Crandall, K.A., Kelsey, C.R., Imanichi, H., Lane, H.C. and Salzman, N.P. (1999) Parallel evolution of drug resistance in HIV: failure of nonsynonymous/synonymous substitution rate ratio to detect selection. *Mol. Biol. Evol.* **16,** 372–382.

Domachowske, J.B., Bonville, C.A., Dyer, K.D. and Rosenberg, H.F. (1998) Evolution of antiviral activity in the ribonuclease A gene superfamily: evidence for a specific interaction between eosinophil-derived neurotoxin (EDN/RNase 2) and respiratory syncytial virus. *Nucleic Acids Res.* **26,** 5327–5332.

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

Dykhuizen, D. and Hartl, D.L. (1980) Selective neutrality of 6PGD allozymes in *E. coli* and the effects of genetic background. *Genetics* **96,** 801–817.

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

Farah, C.S. and Reinach, F.C. (1995) The troponin complex and regulation of muscle contraction. *FASEB J.* **9,** 755–767.

Galtier, N. (2001) Maximum-likelihood phylogenetic analysis under a covarion-like model. *Mol. Biol. Evol.* **18,** 866–873.

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

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

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.

Haydon, D.T., Bastos, A.D., Knowles, N.J. and Samuel, A.R. (2001) Evidence for positive selection in foot-and-mouth-disease virus genes from field isolates. *Genetics* **157,** 151–154.

Hughes, A.L. (1999) *Adaptive Evolution of Genes and Genomes*, Oxford University Press, Oxford, UK.

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

Kimura, M. (1983) *The Neutral Theory of Molecular Evolution*, Cambridge University Press, Cambridge, UK.

Leigh Brown, A.J. (1997) Analysis of HIV–1 *env* gene reveals evidence for a low effective number in the viral population. *Proc. Natl. Acad. Sci. USA* **94,** 1862–1865.

Li, W.-H. (1985) Accelerated evolution following gene duplication and its implications for the neutralist-selectionist controversy. In *Population Genetics and Molecular Evolution* (Eds., Otha, T. and Aoki, K.), Japan Scientific Press, Tokyo, pp. 333–352.

Lynch, M. and Conery, J.S. (2000) The evolutionary fate and consequences of duplicate genes. *Science* **290,** 1151–1155.

Muse, S.V. and Gaut, B.S. (1994) A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with applications to the chloroplast genome. *Mol. Biol. Evol.* **11,** 715–725.

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

Ohno, S. (1970) *Evolution by Gene Duplication*, Springer-Verlag, Berlin.

Ohta, T. (1988a) Further simulation studies on evolution by gene duplication. *Evolution* **42,** 375–386.

Ohta, T. (1988b) Multigene and supergene families. *Oxf. Surv. Evol. Biol.* **5,** 41–65.

Ohta, T. (1993) Pattern of nucleotide substitution in growth hormone-prolactin gene family: a paradigm for evolution by gene duplication. *Genetics* **134:** 1271–1276.

Ohta, T. (1994) Further examples of evolution by gene duplication revealed through DNA sequence comparisons. *Genetics* **138,** 1331–1337.

Plikat, U., Nieselt-Struwe, K. and Meyerhans, A. (1997) Genetic drift can dominate short-term human immunodeficiency virus type 1 *nef* quasispecies evolution in vivo. *J. Virol.* **71,** 4233–4240.

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

Rosenberg, H.F. and Domachowske, J.B. (1999) Eosinophils, riobnucleases and host defence: solving the puzzle. *Immunol. Res.* **20,** 261–274.

Rosenberg, H.F., Dyer, K.D., Tiffany, H.L. and Gonzalez, M. (1995) Rapid evolution of a unique family of primate ribonuclease genes. *Nature Genet.* **10,** 219–223.

Schmidt, T.R., Goodman, M. and Grossman, L.I. (1999) Molecular evolution of the *COX7A* gene family in primates. *Mol. Biol. Evol.* **16,** 619–626.

Sharp, P.M. (1997) In search of molecular Darwinism. *Nature* **385,** 401–404.

Swanson, W.J., Yang, Z., Wolfner, M.F. and Aquadro, C.F. (2001) Positive Darwinian selection in the evolution of mammalian female reproductive proteins. *Proc. Natl. Acad. Sci. USA* **98,** 2509–2514.

Ward, T.J., Honeycutt, R.L. and Derr, J.N. (1997) Nucleotide sequence evolution at the kappa-casein locus: evidence for positive selection within the family Bovidae. *Genetics* **147,** 1863–1872.

Yang, Z. (1997) PAML: a program package for phylogenetic analyses by maximum likelihood. *Cabios* **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. (2001) Adaptive molecular evolution. In *Handbook of Statistical Genetics* (Eds., Balding, D.J., Bishop, M. and Cannings, C.), Wiley & Sons, New York, NY, pp. 327–350.

Yang, Z. and Nielsen, R. (2002) Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. *Mol. Biol. Evol.* **19,** 908–917.

Yang, Z. and Swanson, W.J. (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., Nielsen, R., Goldman, N. and A.-M.K. Pederson, W.J. (2000) Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics* **155,** 431–449.

Zanotto, P.M. de A., Kallas, E.G., de Souza, R.F. and Holmes, E.C. (1999) Genealogical evidence for positive selection in the *nef* gene of HIV–1. *Genetics* **153,** 1077–1089.

Zhang, J., Rosenburg, H.F. and Nei, M. (1998) Positive Darwinian selection after gene duplication in primate ribonuclease genes. *Proc. Natl. Acad. Sci. USA* **95,** 3708–3713.