A Codon-Based Model of Host-Specific Selection in Parasites, with an Application to the Influenza A Virus

Roald Forsberg and Freddy Bugge Christiansen

Parasites are capable of expanding their host range by acquiring a new host species. After a host change event, the selective regime acting on a given parasite gene may change as a result of host-specific adaptive alterations of protein functionality or host-specific immune-mediated selection. We present a codon-based model that attempts to include these effects by allowing the position-specific substitution process to change in conjunction with a host change event. Following maximum-likelihood parameter estimation, we employ an empirical Bayesian procedure to identify candidate sites potentially involved in host-specific adaptation. We discuss the applicability of the model to the more general problem of ascertaining whether the selective regime differs in two groups of related organisms. The utility of the model is illustrated on a data set of nucleoprotein sequences from the influenza A virus obtained from avian and human hosts.

Introduction

Parasite species are capable of expanding their host range. These host radiations may have a detrimental effect on the new host species. Well-known examples include the influenza pandemic of 1918–1920 known as the Spanish Influenza (Gorman et al. 1991), the recent epidemic of the Nipah virus (Chua et al. 2000), and the current epidemic of acquired immunodeficiency syndrome in humans (Gao et al. 1999). A successful host radiation consists of three steps: interspecies transmission to the new host, replication within individuals of the new host, and transmission between individuals of the new host species. The intensity of interspecies transmissions is determined by ecology and behavior and must be quite high among coexisting host species with intimate contact, such as, e.g., predators and their prey. Yet, host radiation events remain rare, and most parasites have a clearly defined and constant host range. The host radiation process must therefore be limited by the latter two steps of replication and transmission. An explanation for this limitation is offered by the general theory of ecological specialization (Turner and Elena 2000). This holds that adaptations to different habitats are antagonistic so that the adaptation of a parasite to replication and transmission in its native host creates a barrier against replication and transmission in other host species. In support of this theory, it has been demonstrated experimentally that parasite adaptation indeed occurs upon the introduction to a new host, and that this happens at the cost of maladaptation to the original host (Crill et al. 2000; Turner and Elena 2000), a phenomenon that has been exploited for decades in the production of live attenuated vaccines through repeated passage on non-host cells (Ebert 1998; Nielsen et al. 2001).

When a host radiation event occurs, it is therefore expected to be accompanied by adaptive evolutionary changes that establish a new phenotype and allow the parasite to complete its life cycle in the new host. Thus, identification of the positions in the parasite genome underlying the phenotypic difference between host-specific strains may provide information about the molecular basis for species-specific adaptation.

Previous studies aimed at this adaptation have focused on identifying fixed species-specific changes in the genomes of parasites from different host species (Hughes et al. 2000; Chang, Sgro, and Parrish 1992). These are natural candidates readily identified from alignment data. However, adaptive alteration of protein functionality may not occur only through fixation of different amino acids in different hosts. It may present itself as a more subtle change in the suite of amino acids that are allowed to occur, resulting not in fixation, but in a different substitution process in the two species. Additionally, host-specific adaptation can occur continuously as a result of host-specific immune selection that continuously selects for protein variants with new antigenic configurations. Fixation of genetic variants is therefore too rigid a criterion for the identification of positions involved in species-specific adaptation. Instead, a comparative methodology which focuses on changes in the evolutionary process in the different host environments may provide a fruitful approach.

To this end, we elaborate a simple codon-based model of nucleotide substitution which describes changes in the selective regime at the protein level that may affect the genes of a parasite confronted with a new host environment. The purpose of the model is twofold: First, we aim to gain a better understanding of the evolutionary processes in parasites that undergo species change. Second, we wish to use the model to identify positions in parasite genomes that may be involved in host-specific adaptations. These may then act as candidates for further experimental work on the molecular biology of species-specific adaptation.

The fundamental idea of the model is to infer how selection acts on variants of parasite genes in different hosts. To do so, we use a site-specific comparison of the substitution process of synonymous (silent) and nonsynonymous (amino acid altering) nucleotide substitutions in the parasite populations of different hosts. This type of comparison has previously been shown to be a useful analytical tool in the study of molecular evolution (Kimura
Nodes of divergence the virus was transmitted from the original host to the new host. A change in the site-specific substitution process of amino acids. Alternatively, positions that are important for protein function in the original host may lose their function in the new host and therefore experience relaxed functional constraints after host radiation. Finally, the new host environment may alter the pattern of immune-mediated selection. The location of antigenic regions may change with ensuing changes in the substitution pattern along the sequence. Also, the epidemiological dynamics may change in the new host, thereby altering the strength of immune-mediated selection. Suppose the dynamics of the parasite in the original host is that of a childhood disease where the parasite most commonly infects immunologically naive individuals. Immune-mediated selection may therefore be limited or absent. Properties of the new host may, however, dictate a different epidemiological dynamic, where the parasite is dependent on continuous reinfection of previously infected individuals and therefore subjected to a strong immune-mediated selection that favors changes in its antigenic properties.

These host-mediated changes in the selective regime will potentially result in an observable host-mediated difference in the position-specific ratio of the two kinds of nucleotide substitutions, and hence will allow for the identification of codon positions involved in host-specific adaptation.

Data

We consider a sample of coding sequences from homologous genes taken from a virus that infects two different species (fig. 1). At some point in its evolutionary history the parasite acquired a new host species. Assuming no recombination, the relationship between the sequences can be described by a tree in which one clade represents sequences evolved in the new host and the rest of the tree represents parasites evolved in the original host. In the tree shown in figure 1 and in the following formulae the root of the tree represents the point of host radiation, but this could equally well be represented by an internal node in the tree. The differentiation between the original host and the new host is used throughout this article for the sake of clarity. It is not necessary, however, to know which species is the original host as long as the point of divergence between the two groups of sequences can be identified in the tree.

Basic Markov Model of Codon Substitution

Suppose the gene sequences in question consist of $n$ codons. The data at site $h (h = 1, \ldots, n)$ are represented by two vectors $o_h$ and $n_h$, where $o_h$ is a vector of codons from the sequences in the original host species at site $h$, and $n_h$ is a vector of codons from the sequences in the new host species at site $h$. The nucleotide-substitution process is described by a codon-based model of substitution, which is a modification of the model proposed by Goldman and Yang (1994). This is given by a $61 \times 61$ matrix (stop codons not allowed) of the relative instantaneous substitution rates, where the rate of transition from codon $i$ to codon $j$ at site $h$ is
The transition probability matrix of codon substitution over a branch in the tree of length \( t \) is then calculated as \( P(t) = e^{\omega t} \). The parameter \( \omega \) modifies the ratio between the rates of nonsynonymous and synonymous substitution and describes the propensity of sites to accept amino acid-altering substitutions. The parameter \( \kappa \) represents the transition/transversion rate ratio, \( \pi_j \) is the equilibrium frequency of codon \( j \) (estimated as the observed codon frequency in the data) and the parameter \( \nu \) is a scaling factor defined by the requirement that the average rate of substitution be one:

\[
\sum_{\forall (i,j), \forall \phi} \pi_j q_{ij} = 1.
\]

This scaling means that branch lengths can be interpreted as the expected number of nucleotide substitutions per codon averaged over all sites.

Allowing for Heterogeneous Selection Pressure

The selective regime may differ at different positions in the gene, and potentially in different hosts. To reflect this difference, we allow a statistical distribution \( p(\omega) \) of \( \omega \) ratios among sites (Yang et al. 2000). The probability of observing the data in a site is then obtained by integration,

\[
P(data) = \int p(\omega) P(data | \omega) d\omega.
\]

Constant Selection Pressure After Host Radiation

Consider a codon in the alignment. We assume that the selective regime is constant in the clade that represents viruses from the original host. The simplest event that may occur at the time of host radiation is the event \( c \), that the selective regime in this codon position (and therefore \( \omega \)) remains the same. Conditional on event \( c \) we have that the total probability of the data in site \( h \) is given by

\[
P(o_h, n_h | c, M) = \sum_{j=1}^{61} \pi_j \int P(x_j, o_h | M, \omega) P(x_j, n_h | M, \omega) p(\omega) d\omega.
\]

Here, \( x_j \) is the codon state in the node of divergence (fig. 1), \( M \) represents the model parameters \( \kappa, \pi, \) and the tree, including topology and branch lengths. The probability of observing codon \( x_j \) given the subtrees from the two different species, \( P(x_j, o_h | M, \omega), P(x_j, n_h | M, \omega) \), is found by traversing the subtrees according to Felsenstein’s pruning algorithm (Felsenstein 1981). We can now construct the simplest hypothesis, \( H_1 \), that no changes occur after host radiation. The probability of the data under this hypothesis is simply

\[
P_h(H_1 | M) = P(o_h, n_h | c, M),
\]

and under the assumption that sites evolve independently, the full likelihood of the data is

\[
L_{H_1} = \prod_h P_h(H_1 | M).
\]

Describing Host-Mediated Changes in the Selective Regime

To allow for the selective regime to change after the introduction to the new host, we include in our model the event \( d \), which specifies a divergence in the selective regime affecting a site after host radiation. This means that if \( d \) occurs, then the ratio of the rate of nonsynonymous to synonymous substitutions \( (\omega) \) on the branches of the tree that represent evolution in the new host is permitted to differ from that in the original host. Lacking any prior assumptions concerning the direction of the change, we simply say that if \( d \) occurs, then the choices of the \( \omega \) parameters in the two hosts become independent of each other. Conditional on \( d \), the probability of observing \( o_h \) and \( n_h \) is therefore obtained by independently integrating over the \( \omega \) distribution in the two parts of the tree and multiplying

\[
P(o_h, n_h | d, M) = \sum_{j=1}^{61} \pi_j \int P(x_j, o_h | M, \omega) p(\omega) d\omega 
	\times \int P(x_j, n_h | M, \omega) p(\omega) d\omega.
\]

However, the selective regime on all sites in a protein is not likely to change after host radiation. Therefore, in the construction of a hypothesis \( H_0 \) that allows for host-specific selection we let the event \( d \) occur for a site with probability \( p_d \). We then have that the total probability of observing the data under \( H_0 \) is

\[
P_h(H_0 | M) = p_d P(o_h, n_h | d, M)
	+ (1 - p_d) P(o_h, n_h | c, M),
\]

where \( p_d \) is a parameter that describes the propensity of a site to experience a change in the selective regime. The total likelihood is found by a similar expression to (1).

Comparing Models by Likelihood Ratio Tests

The two hypotheses are nested in that \( H_1 \subset H_0 \). Comparison of the models, however, corresponds to fixing the parameter \( p_d \) at the boundary. In this situation, evaluation of the test statistic using the \( \chi^2 \) distribution will provide a conservative test of significance. To avoid this effect, the simple \( \chi^2 \) distribution can be replaced with a mixed distribution when comparing test statistics that are close to the critical value (Ota et al. 2000).

Identification of Candidate Sites

To identify sites where the selective regime has changed after host radiation, we use an empirical Bayes approach (Nielsen and Yang 1998; Yang et al. 2000).
Following maximum likelihood estimation of the parameters in the model \((M, \hat{\rho}_d)\), we can calculate the probability of event \(d\) in a given site \(h\) as

\[
P_h(d \mid o_h, n_h, \hat{M}) = \frac{\hat{\rho}_d P(o_h, n_h \mid d, \hat{M})}{\hat{\rho}_d P(o_h, n_h \mid d, \hat{M}) + (1 - \hat{\rho}_d)(P(o_h, n_h \mid c, \hat{M})}.\]

**Implementation and Results**

A computer program written in JAVA that implements these models was constructed using parts of the PAL library (Drummond and Strimmer 2001) and can be downloaded from the Web site http://birc.dk.

To allow for heterogeneous selection regimes, we first implemented a discrete version of a continuous Gamma distribution following Yang et al. (2000). However, as have others, we encountered several problems with this approach, and it was therefore abandoned in favor of a discrete distribution (Yang et al. 2000). Say we have \(K\) classes of sites in the alignment with the \(\omega\) ratios and proportions given as

\[
p_0 p_1 \cdots p_{K-1} \quad \omega_0 \omega_1 \cdots \omega_{K-1}.
\]

This adds \(K\) rate ratio and \(K - 1\) class-probability parameters to the model.

To demonstrate the utility of our model, we analyzed a data set of nucleoprotein (NP) sequences from the influenza A virus taken from avian and human hosts. The NP gene encodes a protein with 498 amino acids whose primary function is the binding of viral RNA segments to form the nucleocapsid of the virus particle (Lamb and Krug 1996). The NP gene has been assigned a putative role as a determinant of host range (Murphy and Webster 1996), but the identification of residues important for species-specific adaptation has been hindered by the lack of a resolved three-dimensional protein structure (Lamb and Krug 1996). Molecular biological analysis, however, has revealed regions of the NP that are of functional importance and several epitopes that are supposedly recognized by cytotoxic T lymphocytes (CTL) of the cellular immune response (Neumann, Castrucci, and Kawaoka 1997; DiBrino et al. 1993; Davey, Dimmock, and Colman 1985; Voeten et al. 2000; Boon et al. 2002). The present-day form of the human influenza. A nucleoprotein is believed to have entered the human population as an avian virus was transferred to humans immediately before the onset of the Spanish Influenza pandemic of 1918–1920 (Gorman et al. 1991). Since then, the nucleoprotein gene has evolved independently in human and avian hosts. In contrast, other genetic elements were exchanged between avian and human influenza viruses prior to each of the two pandemics in the last half of the twentieth century (Murphy and Webster 1996).

Sequences were downloaded from the Influenza Sequence Database (Macken et al. 2001) and include all available avian NP sequences and an equal number of human NP sequences that were chosen to cover the largest possible time span. Figure 2 shows the genealogy of the chosen isolates. This was inferred under the maximum likelihood (ML) criteria by the fastDNAml algorithm (Olsen et al. 1994) provided at http://bioweb.pasteur.fr/seqanal/interfaces/fastdnaml-simple.htm. A very similar genealogy was inferred with the Neighbor-Joining (NJ) algorithm as implemented in the PAUP software (Swofford 2002). Furthermore, a Bayesian estimate of the posterior genealogy distribution was performed using the MrBayes program (Huelsenbeck and Ronquist 2001) with a general time reversible (GTR) model of substitution and a gamma distribution on rate heterogeneity. The estimation involved 3 million Markov Chain Monte Carlo samples and showed that besides the ordering of a few short external branches, there was high confidence in the structure of the ML tree (results not shown). The position of the root was determined by the inclusion of an equine influenza virus (Gorman et al. 1991), and because this isolates the human influenza lineages as a monophyletic group, the root was used as the time of the transmission event. Branch lengths were estimated using the CODEML program of the PAML package under a model that allows three discrete classes of \(\omega\) rates (Yang 2000).

Analyses were performed using the genealogies estimated by ML and NJ. This yielded very similar parameter estimates and site-specific probabilities, indicating that the methodology is robust to minor differences in the estimated genealogy. For this reason we have listed only results obtained using the ML tree. Results of the analysis under the different hypotheses, using three classes of \(\omega\) ratios, are shown in table 1. Hypothesis \(H_0\) allows the ratio between the rates of nonsynonymous to synonymous substitution (\(\omega\)) to differ in the two host species, and it provides a significantly better fit to the data than \(H_1\), which does not include this feature \((P_{\chi^2} < 0.01)\). It is also evident from table 1 that the estimated distribution of \(\omega\) values is wider under \(H_0\) than under \(H_1\), and that it also encompasses a class of positively selected sites (\(\omega > 1\)). The empirical Bayes assignment of site-specific probabilities of \(\omega\) divergence identified 12 sites with a high probability \((P > 0.90)\) of \(\omega\) divergence. These are listed in table 2, which also lists molecular biological results on NP function and immunogenicity. One of these sites (334) is located in a region proposed to be involved in the transport of NP over the nuclear membrane; another site (423) is located in a proposed CTL epitope; and two sites (32, 343) are located in regions that are proposed to be involved in both nuclear transport and CTL recognition. Furthermore, two sites (350, 353) are located in the flanking region of a proposed CTL epitope where mutations are known to disrupt cytotoxic processing of viral CTL epitopes. Six sites, however, are located in regions of NP where no biological role has yet been proposed. Interestingly, two of these (101, 102) are neighbors, whereas the remaining four are spaced further apart (62, 131, 214, 290).

**Discussion**

The application of our model to the NP data showed that the hypothesis, which allows the position-specific substitution process to differ between parasites occupying
different hosts, provides the better description of data. We believe that our model captures a relevant biological phenomenon in the evolution of this host-changing parasite: namely, that selection favors parasite proteins with different properties in different hosts and therefore creates different patterns of evolution. Furthermore, we were encouraged to find that several of the sites identified by our method correspond to regions of the NP gene known to be of functional or immunological importance in the human host. An additional six sites were identified in regions of unknown function and constitute interesting targets for molecular biological analysis, such as directed...
multiplying all sequences from the new (human) host. This was done by attempt to include this possibility, we also implemented an excess of nonsynonymous substitutions in the time we would, however, expect these adaptive changes to create its signature in the ensuing evolutionary process. We inferring the adaptive evolutionary change itself, but rather tive move in sequence space. In the latter case, we are not the selective regime has changed, either as a result of adaptation in the influenza virus.

mutagenesis, to evaluate their effect on host-specific adaptation in the influenza virus.

This approach focuses on identifying sites where the selective regime has changed, either as a result of host-specific alterations in immune-mediated selection or as a result of altered functional constraints after an adaptive move in sequence space. In the latter case, we are not inferring the adaptive evolutionary change itself, but rather its signature in the ensuing evolutionary process. We would, however, expect these adaptive changes to create an excess of nonsynonymous substitutions in the time immediately following the host radiation event. In an attempt to include this possibility, we also implemented a model that allows for a burst of nonsynonymous substitutions on the uppermost branch of the subtree relating the sequences from the new (human) host. This was done by multiplying all $\omega$ values in the distribution by a positive number ($f_\beta$) when calculating the substitution probabilities on this branch, thus increasing their value. The results of this series of calculations were, however, not easily interpretable, as the model seemed to favor lower $\omega$ values on this branch ($f_\beta < 1$), contrary to our expectation.

The inference of parameters like $f_\beta$ that have their effect deep within an evolutionary tree is an intrinsically difficult problem because the uncertainty of the states at the inner nodes of a tree increases as one progresses upward from the leaves, making it difficult to determine whether the above-mentioned result is an actual phenomenon, or an artifact of the methodology. A potential explanation is, that the chosen node of divergence is a bad estimate of the true time of divergence. A branch in a tree is an arbitrary evolutionary unit affected by sampling effects. Hence, the inclusion of further avian lineages might break up the branch leading to the clade of human influenza viruses, signifying that part of the evolution on this branch occurred in the original avian hosts rather than in the new human host. In this case, a generally higher rate of nonsynonymous substitution on the human part of this branch would give a poor description of the evolutionary regime predominant in the true avian host. A way to counter this effect would be to lower the $f_\beta$ parameter and thus the number of nonsynonymous substitutions.

There are two interesting implications of this that we did not pursue. It may be possible to infer the maximum likelihood position of the species transmission event along this branch, which in conjuction with a molecular clock model could provide an estimated date of species transmission. Furthermore, it may be possible to infer which host species is most likely to be new to the parasite if the history of the parasite is unknown.

The motivation for developing the model presented here was to study the evolution of parasites undergoing host radiation, but the model addresses the more general problem of describing the substitution process in two groups of related organisms and determining whether and where the selective regime may differ. In our formulation, the evolutionary event that is believed to have changed the way natural selection acts is the introduction to a new host environment. A closely related problem, and one that has received considerable attention, is that of functional divergence in paralogous genes, where the decisive evolutionary event is a duplication of the genetic element.

It is therefore not surprising that two recent models of functional divergence in paralogous genes are conceptually similar to our model in using models of amino acid substitution that allow the position-specific rate of amino acid substitution to change after gene duplication (Gu 2001; Knudsen and Miyamoto 2001). However, these models infer changes in the selective regime through a change in the absolute rate of substitution, whereas the codon-based approach presented here uses a comparison between the two different types of nucleotide substitution, thus enabling use of the full information in the nucleotide sequence and inclusion of the known biological phenomenon of transition-transversion bias. A potential advantage of the amino acid-based models is that they use empirically based transition matrices, which may include the effect that the physiochemical properties of amino acids have on the substitution process.

To simplify our model, we have not included any such empirical priors on substitution rates, but such factors could easily be included in the codon-based substitution model (see Goldman and Yang 1994). In relation to this aspect of the model, the $\omega$ parameter used is a rather crude indicator of the selective regime, and it would be interesting to consider a model which allowed additional

### Table 1

<table>
<thead>
<tr>
<th>Hypothesis</th>
<th>Likelihood Value</th>
<th>Parameter Estimates</th>
</tr>
</thead>
<tbody>
<tr>
<td>$H_1$</td>
<td>$-12824.26$</td>
<td>$\omega_0 = 0.007$, $\omega_1 = 0.12$, $\omega_2 = 0.78$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$p_0 = 0.68$, $p_1 = 0.26$, $p_2 = 0.06$</td>
</tr>
<tr>
<td>$H_0$</td>
<td>$-12799.65$</td>
<td>$\omega_0 = 0.005$, $\omega_1 = 0.14$, $\omega_2 = 4.03$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$p_0 = 0.71$, $p_1 = 0.24$, $p_2 = 0.05$</td>
</tr>
</tbody>
</table>

### Table 2

<table>
<thead>
<tr>
<th>Codon position ($P &gt; 0.90$)</th>
<th>P</th>
<th>Proposed Function of Region</th>
</tr>
</thead>
<tbody>
<tr>
<td>33</td>
<td>0.98</td>
<td>1–38 nuclear export signal</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(Neumann et al. 1997)</td>
</tr>
<tr>
<td>32–40 CTL epitope</td>
<td></td>
<td>(DiBrino et al. 1993)</td>
</tr>
<tr>
<td>63</td>
<td>0.91</td>
<td>None</td>
</tr>
<tr>
<td>101</td>
<td>0.94</td>
<td>None</td>
</tr>
<tr>
<td>102</td>
<td>0.99</td>
<td>None</td>
</tr>
<tr>
<td>131</td>
<td>0.99</td>
<td>None</td>
</tr>
<tr>
<td>214</td>
<td>0.92</td>
<td>None</td>
</tr>
<tr>
<td>290</td>
<td>0.99</td>
<td>None</td>
</tr>
<tr>
<td>334</td>
<td>0.96</td>
<td>327–345 nuclear membrane transport (Davey et al. 1985)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(Voeten et al. 2000)</td>
</tr>
<tr>
<td>343</td>
<td>0.96</td>
<td>338–347 CTL epitope</td>
</tr>
<tr>
<td>350</td>
<td>0.93</td>
<td>Flanking epitope above</td>
</tr>
<tr>
<td>353</td>
<td>0.94</td>
<td>Flanking epitope above</td>
</tr>
<tr>
<td>423</td>
<td>0.92</td>
<td>418–426 CTL epitope</td>
</tr>
</tbody>
</table>

**Note.**—P, posterior probability of divergence in the given codon position.
indicators such as, e.g., shifts between biochemically different groups of amino acids.

There are additional ways in which the present approach could be improved. An apparent and readily made generalization, is to extend the model to allow for several independent host change events—i.e., a tree in which several clades of the new host are embedded in a larger tree of isolates from the original host. In the construction of the model, we have made the implicit assumption that the equilibrium codon frequencies do not change after host change. The validity of this assumption is not clear, but changes in the host cell environment could potentially change the direction of any codon bias. A simple way to include this assumption could be to estimate the equilibrium codon frequencies from the two groups separately and then apply separate transition matrices to the different parts of the tree.

Another assumption is that the distribution of $\omega$ is the same in both species—i.e., that the change in $\omega$ is symmetric so that an upward change in one site is balanced by a downward change in another site. This assumption is justified by the recognition that there is limited information in the data to estimate two independent distributions and that any bias it induces on the identification procedure will be toward more conservative estimates. A more grave assumption is that of independence between codon positions. This is a commonly made assumption in molecular evolutionary analysis of proteins, but it is inherently wrong because of the known three-dimensional nature of biological molecules. Some efforts have been made to address this issue (Pedersen and Jensen 2001; Pollock, Taylor, and Goldman 1999), but both approaches are computationally very demanding and were therefore not pursued here.

Rates of mutation may change after a viral host change because of alterations in, e.g., the rate of replication. As stated previously, we have assumed that the rates of synonymous and nonsynonymous substitution scale identically with such changes in the mutation rate. The validity of this assumption depends on the unknown population genetic mechanisms underlying the process of fixation. For sites where all amino acid substitutions are either neutral or strongly deleterious, we expect this scaling to hold. This may not be the case however, for sites undergoing positive selection if the process of fixation is limited by factors other than the availability of mutations. Should this assumption be violated, a change in the mutation rate could cause a change in the ratio of the rates of synonymous and nonsynonymous substitution to occur without there being any change in the underlying selective regime acting at the protein level, and therefore, bias analyses like the one presented here lead to the identification of false positives. In our future work, we intend to explore and potentially eliminate this bias through the construction of a model which allows for simultaneous changes in both the selective regime acting at the protein level and the underlying rate of mutation as represented by the rate of synonymous substitution.

All this having been said, we still believe that the results from the analysis of the NP gene data show promise for our comparative approach and that this type of analysis, correlated with structural, immunological, and other information, may provide a useful way to identify candidate sites for the molecular biological investigation of species-specific adaptation in parasites.

Acknowledgments

We thank Jakob Skou Pedersen, Mikkel Heide Schierup, Bjarne Knudsen, and Alexei Drummond for useful comments and discussions. This study was supported by grants 21-02-0206 and 51-00-0392 from the Danish Natural Science Research Council; by grant 1R01-GM60729-01 from the National Institutes of Health, USA; by grant HAMJW from EPSRC; and by grant HAMKA from MRC.

Literature Cited


