Detecting Adaptive Molecular Evolution: Additional Tools for the Parasitologist

James O. McInerney[1], D. Timothy J. Littlewood[2] and Christopher J. Creevey[1]

[1]Bioinformatics and Pharmacogenomics Laboratory, Department of Biology, National University of Ireland, Maynooth, Co. Kildare, Ireland
[2] Parasitic Worms Division, Department of Zoology, The Natural History Museum, Cromwell Road, London SW7 5BD, England, UK

Abstract 360
1. What is Adaptive Molecular Evolution? 360
2. Methodological Advances 362
   2.1. Pair-wise Comparisons 362
   2.2. Phylogeny-Based Methods 363
3. Example of Adaptive Evolution in the Malaria RIFIN Proteins 366
   3.1. Methods 367
   3.2. Results 368
      3.2.1. Pair-wise Distance Method 368
      3.2.2. Messier and Stewart Method 368
      3.2.3. Relative Rate Ratio Test (RRRT) 370
      3.2.4. Likelihood Ratio Test (LRT) of positive Selection 370
      3.2.5. Bayesian Inference of Codons under Positive Selection 373
4. Prospects 373
Acknowledgements 376
References 377

ADVANCES IN PARASITOLOGY VOL 54
0065-308X $35.00

ABSTRACT

It is likely that infectious diseases have shaped the evolution of many vertebrates, including humans. The etiological agents of disease continuously strive to evade the immune response and the immune response, in turn, seeks to change in order to keep pace with the invaders. This 'arms race' may be characterized by the selection for new variant hosts and new variant parasites. Here we discuss the utility of phylogenetics in detecting adaptive evolution at the molecular level and, for illustration, we concentrate on a family of surface-exposed proteins (the rifins) found in the recently sequenced genome of *Plasmodium falciparum*. We employed phylogeny-based methods in order to characterize adaptive evolution in these proteins. We found evidence for adaptive evolution in many of the amino acid residues in at least one lineage. These results indicate that there has been selection for those strains of *P. falciparum* that contain the new genotypes. These proteins are likely to be of great importance for the survival of the parasite. Studies of the interaction of these proteins with the antigen-presenting cells of the immune system should lead to a better understanding of malarial infection.

1. WHAT IS ADAPTIVE MOLECULAR EVOLUTION?

Proteins are exquisitely designed molecular machines that have been honed and perfected by a system of mutation and selection over time. Mutation relentlessly increases variation which selection edits, either removing (as in negative selection) or accepting it (as in positive directional or non-directional selection). For that reason, selection is likely to have been a powerful agent of evolution, largely dictating the rate and pattern of protein change. Usually, we think of the main role of selection in terms of rejection of new variants that are considered deleterious. However, it is becoming more obvious that adaptive change (positive selection of new variants) is a frequent and essential component of molecular evolution (McDonald and Kreitman, 1991; Swanson and Vacquier, 1995; Messier and Stewart, 1997; Nielsen and Yang, 1998; Zanotto *et al*., 1999; Bielawski and Yang, 2001).

If a new mutation occurs, a number of factors will influence whether or not this mutation remains polymorphic in the population for a long period. If the new mutation is neither significantly advantageous nor disadvantageous, then the length of time it remains in the population as a polymorphism along with the wild type is dependent on the effective population size of the species. If the new mutation affects the phenotype then it either

may be removed or become fixed in the population in a manner that is almost independent of the population size depending whether the mutation is under positive or negative selection (Hughes, 1999).

For the most part, the effects of selection are analysed in protein-coding regions. Protein-coding genes are a mosaic of nucleotide positions and mutations within them can be classified as either synonymous or nonsynonymous, depending on whether or not the mutation at that codon site changes the encoded amino acid. The most commonly used method of detecting adaptive evolution at the protein level usually has involved comparing rates of synonymous changes per synonymous site (d_s) versus nonsynonymous changes per nonsynonymous site (d_n). The synonymous rate is used to calibrate the properties of neutrally evolving DNA so that when it is compared against the nonsynonymous rate the action of negative or positive selection may be detected. Negative selection, being the retention of status quo, is characterized by a synonymous rate that is much greater than the nonsynonymous rate. Positive selection, being the advocate of change, is characterized by a nonsynonymous rate that is much greater than the synonymous rate.

It is usual to think of adaptive evolution as an ephemeral event that is associated with the development of a new function or the modification of an existing property of a protein. A period of positive selection might be supplanted by a period of purifying selection once this novel property has been established. The transient nature of this phenomenon, allied to its enormous potential for affecting phenotypic diversity, makes it important to understand its nature and extent.

For example, it has been reported that periods of adaptive evolution have characterized the development of new functions in the lysozyme proteins of foregut fermenting primates (McInerney, 1998; Yang, 1998; Creevey and McInerney, 2002). In this situation, a lysozyme protein that is active in low pH conditions and is unusually resistant to cleavage by pepsin is found in the fermenting foregut of herbivorous primates. The unusual properties of lysozyme allow the digestion of plant material that would otherwise have remained refractory. In some marine invertebrates, positive selection for new variant sperm proteins has also been reported and implicated to be responsible for the origin of new species (Yang *et al*., 2000a). The biological 'arms race' that characterizes the interaction between pathogenic microorganisms and their vertebrate hosts has also been documented in terms of adaptive evolution of important components of both systems (Hughes, 1992; Hughes and Hughes, 1995).

In a seemingly contradictory finding, the importance of any individual gene for the survival of a species may be relatively small. In yeast, it has been shown that deleting a gene has little phenotypic effect owing to the existence

of compensatory mechanisms (Winzeler *et al*., 2000). This could indicate that individual residues within genes are even more dispensable (and presumably not under a great deal of selective pressure). However, running contrary to this expectation are the observations of selective pressures on most genes (Creighton and Darby, 1989; Hughes, 1992; McInerney, 1998; Bielawski and Yang, 2001).

2. METHODOLOGICAL ADVANCES

Although it is becoming increasingly obvious that the evolutionary histories of many genes have been characterized by intermittent periods of adaptive evolution, it is still a controversial and difficult task to correctly identify those periods and to identify the amino acids that were influenced by positive selection (Yang and Bielawski, 2000). Initially, the most commonly used methods relied on simple pair-wise comparisons of protein-coding sequences (Li *et al*., 1985; Nei and Gojobori, 1986; Li, 1993; Ina, 1995). More recently, however, phylogenetic trees have been employed in order to pinpoint adaptive evolutionary events with greater accuracy (Messier and Stewart, 1997; Yang, 2000; Yang and Bielawski, 2000; Creevey and McInerney, 2002; Yang and Nielsen, 2002).

2.1. Pair-wise Comparisons

One of the first pair-wise distance methods was devised by Li *et al*. (1985). This method was later revised (Li, 1993) and other modifications of this principle have also been described (Ina, 1995). According to Nei and Gojobori (1986), for each pair of protein-coding sequences in an alignment, each codon position is classified according to whether or not there has been a change since the two sequences last shared a common ancestor. These variable positions are classified as synonymous (silent) or nonsynonymous (replacement) substitutions. Two pair-wise 'distances' are calculated – the proportion of synonymous substitutions per synonymous site (pS) and the proportion of nonsynonymous substitutions per nonsynonymous site (pN):

$$p\mathrm{S} = \mathrm{Sd}/\mathrm{S} \tag{2}$$

$$p\mathrm{N} = \mathrm{Nd}/\mathrm{N} \tag{3}$$

where Sd is the number of synonymous differences and S is the number of synonymously variable sites and Nd is the number of nonsynonymous

differences and N is the number of nonsynonymous sites. A log-normal correction (Jukes and Cantor, 1996) for superimposed substitutions modifies the observed distance and produces an estimate of dN and dS that is larger than the observed distance:

$$d\mathrm{S} = -3\log_e(1-4p\mathrm{S}/3)/4 \quad (4)$$

$$d\mathrm{N} = -3\log_e(1-4p\mathrm{N}/3)/4 \quad (5)$$

These corrected distances are compared and, in those situations where dN is significantly greater than dS, a period of adaptive evolution is invoked as the reason for this occurrence. There are many variations of methods for calculating these distances (Li *et al.*, 1985; Li, 1993; Ina, 1995), but the principle remains the same.

This kind of approach works very well for closely related sequences. The distances can be compared statistically as the variance can be computed and differences that may be due to estimation error can be distinguished from differences that are due to selection.

There are, however, a number of problems with pair-wise distance analyses. First, the rate of change at silent sites is usually quite high. Since substitutions at these sites are either not deleterious or only very slightly deleterious (there may be small selective pressures for optimal codon usage; McInerney, 1998), the silent substitution rate is very close to the mutation rate (Kimura, 1968). Mutation rates have been measured to be of the order of 10^{-9} substitutions per site per year (Gaut and Clegg, 1991, 1993). This means that, while amino acid conservation may be high in some proteins, a considerable amount of change may have occurred at silent positions. This results in a tendency to underestimate the value of dS. The value of dN is not masked to the same extent, although superimposed substitutions also occur at amino acid replacement sites. Additionally, the overall pair-wise distance is an average of the rate of change at all positions. Therefore, if selective pressures are heterogeneous across the protein, this variability is masked. Even in those situations where pair-wise sequence analysis indicates that positive selection has occurred, it is not possible to say on which lineage it has occurred.

2.2. Phylogeny-Based Methods

New methods for detection of adaptive evolution are based on phylogenetic trees. This has been the case since the seminal contribution by Messier and Stewart (1997). Their innovation involved the use of a phylogenetic tree to try to pinpoint the time when an adaptive evolutionary event took place.

Using a dataset of primate lysozyme sequences, Messier and Stewart reconstructed the hypothetical ancestral sequences at each of the internal nodes of the phylogenetic tree. They then performed all possible pair-wise comparisons between all sequences, both hypothetical and contemporary, and found that this provided a greater power when trying to pinpoint the time of adaptive evolutionary events.

There are a number of advantages to this approach. The first stems from the use of both a phylogenetic tree and reconstructed ancestral sequences. Given a phylogenetic tree, it is possible to use either maximum parsimony (MP) or maximum likelihood (ML) methods (Felsenstein, 1981, 1996; Yang *et al*., 1995) to reconstruct the hypothetical sequences at internal nodes. This means that a particular dataset has $2n-2$ taxa: n terminal lineages and $n-2$ internal nodes for a bifurcating unrooted phylogeny. These hypothetical ancestral sequences are distributed throughout the evolutionary history of the analysed sequences, reducing the gap between sampling times. This means that if selective pressures differ across the tree, the increased sampling should allow a greater power in pinpointing when the period of adaptive evolution occurred.

According to the neutral theory, the rate of synonymous to nonsynymous mutations should be the same for intra- and interspecific comparisons of sequences. Initially proposed by Maynard-Smith (1970), this was implemented by McDonald and Kreitman (1991) to detect adaptive evolutionary events in the alcohol dehydrogenase locus of *Drosophila* species. An increase in the number of intraspecific nonsyonymous mutations compared to interspecific nonsynonymous mutations indicates that the intraspecific substitutions were favoured by selection within that species. This approach also implicitly uses phylogenetic trees, albeit simple ones with a single branch separating two closely related species.

Creevey and McInerney (2002) suggested a method that is loosely based on the McDonald and Kreitman test. In this case, a phylogenetic tree is inferred and presumed to be correct. The ancestral sequences at each internal node on this tree are reconstructed. From each internal node, all the substitutions in the descendent clade are characterized and their numbers are counted. Substitutions are classified into replacement-invariable (RI), replacement-variable (RV), silent-invariable (SI) and silent-variable (SV) sites. Replacement-invariable sites are those where a replacement substitution has occurred at a particular codon position somewhere in the clade, but this new amino acid did not subsequently change. Replacement-variable sites are those where a replacement substitution occurred in the descendent clade and, subsequently, this amino acid position changed again. Similar statistics, SI and SV, are calculated for silent substitutions. As with the McDonald–Kreitman test, it is assumed that, in cases where the sequences

are evolving neutrally, there should be no difference between the ratios RI : RV and SI : SV. Once again, the silent substitutions are considered to be neutrally evolving and are used to calibrate the properties of neutrally evolving DNA. In this case, the ratio of SI to SV substitutions in any clade should be determined by the size and shape of the clade and the mutation rate. If the internal branch that describes the clade is close to the tips of the tree, then SI sites will dominate (and indeed, so should RI sites if the sequences are evolving neutrally). If the internal branch is further from the tips, then the ratio of SI to SV will decrease. If positive selection has characterized the evolution of the sequences being examined, the number of either RI or RV substitutions will be significantly greater than is expected from neutrality. The former indicates the presence of positive directional selection, and the latter indicates positive nondirectional selection. This method has been shown to be very effective at identifying adaptive evolution in a number of instances (Creevey and McInerney, 2002).

Maximum likelihood (ML) and Bayesian methods have also been implemented for detection of adaptive evolution (Goldman and Yang, 1994; Yang, 1998; Yang and Nielsen, 2002). Maximum likelihood methods choose a hypothesis that maximizes the likelihood of observing the data. This hypothesis is usually composed of two parts – the tree and the evolutionary process, each with a certain set of parameters. More formally:

$$\mathrm{L} = \mathrm{P}(\mathbf{X}|\tau, \upsilon, \theta) \tag{10}$$

where L is the likelihood and P is the probability of observing $\mathbf{X}$, the alignment given the tree τ with branch lengths υ and the substitution process described by θ.

The substitution process can be described by a continuous time Markov process with transition/transversion rate bias and codon usage bias allowed to vary. In addition, the physicochemical distances between amino acids being is used to accommodate selective restraints at the protein level. Given that selective pressures are likely to vary across different sites in a protein sequence, models have been developed to incorporate heterogeneous selective pressures at different sites (Yang *et al.*, 2000b). Variation in selective pressure is usually modelled according to a statistical distribution or a mixture of statistical distributions (Yang *et al.*, 2000b). As a result, a site has a probability of belonging to a particular class. In addition, the process may vary over the tree. To account for this, mixed models have been developed that allow the dN : dS ratio to vary either among lineages or among sites. In this case, classes of sites may differ between 'foreground' lineages (pre-specified lineages of interest) and 'background' lineages (the other lineages in the tree).

The likelihood ratio test (LRT) is used to evaluate nested models of sequence evolution. Some models are more parameter-rich extensions of other models and, when this is the case, an LRT may be performed with twice the log-likelihood difference being compared with a χ^2 distribution with the degrees of freedom equal to the difference in the number of parameters between the two models (for a more complete description, see Yang *et al*., 2000b).

Finally, when after the ML optimization of the parameters is complete, an empirical Bayes approach is used to infer which class a site is most likely to come from (Nielsen and Yang, 1998). Those sites with a high probability of coming from a class of sites with a high dN : dS ratio are most likely to be under positive selection.

3. EXAMPLE OF ADAPTIVE EVOLUTION IN THE MALARIA RIFIN PROTEINS

We examine the RIFIN family of proteins in the newly sequenced genome of the malarial parasite *Plasmodium falciparum* to demonstrate some of the methods of detecting adaptive evolution mentioned previously. Malaria is a major contributor to mortality and morbidity in developing countries. Its economic impact alone is sufficient to cripple the economy of most countries in which it is endemic (World Health Organization, 2002). Globally, it is estimated that 300–500 million people are infected with *Plasmodium* annually, with an estimated 1.5 to 2.7 million people dying each year from the disease, mostly children under the age of five years (World Health Organization, 1997).

Immunity to malaria usually develops later in life, with individuals usually not acquiring the broad repertoire of antibodies necessary to suppress the disease until they are older (Abdel-Latif *et al*., 2002). The acquired immune system mounts a response to the polymorphic antigens that are expressed on the surface of infected erythrocytes during the blood stage of the *Plasmodium* life cycle (Fernandez *et al*., 1999; Good and Doolan, 1999). These antigens usually display both temporal (Kyes *et al*., 2000) and sequence (Forsyth *et al*., 1989) variation. The best-characterized *P. falciparum* surface antigen is the erythrocyte membrane protein 1 (PfEMP-1), which is encoded by about 50 *var* genes (Abdel-Latif *et al*., 2002). Malaria infection is linked to parasite-induced surface changes of infected red blood cells. These lead to sequestration of infected red blood cells in the microvasculature and rosetting of uninfected red blood cells, with consequent obstruction to microvascular blood flow, tissue damage and disease (Kyes *et al*., 1999). The PfEMP-1 family of proteins mediate

adhesion to at least some host cell receptors, via CD36 with endothelial cells and via complement receptor 1 with uninfected red blood cells (Kyes *et al.*, 1999). PfEMP-1 is, therefore, considered a major virulence factor, but it alone is not sufficient to explain all parasite-induced surface phenotype changes. Recent analysis of the *P. falciparum* genome identified other multicopy gene families unique to the genus, the largest of which belongs to the *rif* (repetitive interspersed family) gene family (Weber, 1988; Kyes *et al.*, 1999). It is estimated that there are in excess of 200 *rif* genes per haploid genome making them at least four times as abundant as the *var* genes. *Rif* genes are located in close association with *var* genes in clusters within 50 kb of the telomeres, and the name rifins has been proposed for the putative protein products (Gardner *et al.*, 1998).

Rif genes possess two exons, the first of which encodes a putative signal peptide and the second of which encodes an extracellular domain made up of a conserved and a variable region, followed by a transmembrane segment and a short intracellular portion (Abdel-Latif *et al.*, 2002). Rifin proteins have been shown to express on the surface of infected red blood cells. While they have a function in rosette formation along with *var* genes, it is not thought to be their primary function, as some parasites that do not show appreciable levels of rosetting still express rifin proteins (Kyes *et al.*, 1999). Unlike *var* genes, only one of which is expressed at any one time, several *rif* genes are believed to be concomitantly expressed on the surface of infected erythrocytes. The function of these clonally variant rifins on the surface of IE remains speculative but, because of their surface locality and sequence diversity, these proteins may play an essential role in the host–parasite interface during the asexual blood stage (Abdel-Latif *et al.*, 2002). Supporting this hypothesis, Abdel-Latif *et al.* (2002) demonstrated a naturally acquired antibody response in a large number of hosts to a subset of erythrocyte surface-localized *rifin* proteins. However, to date the functional role of rifins as a multigene family of clonally variant surface proteins has not been elucidated, but it remains plausible that anti-rifin antibodies could interfere with some important aspect of the life or cell function of the parasite, such as cytoadherence or rosette formation.

3.1. Methods

A total of 51 sequences were extracted from the *P. falciparum* database available at http://www.plasmodb.org/. Members of the family were identified using a BLASTP search (Altschul *et al.*, 1997) using various candidate members of the RIFIN family as query sequence against a protein-coding version of the annotated *Plasmodium* genome. The DNA

version of all sequences was retrieved, conceptually translated into the corresponding protein, aligned using the ClustalW alignment software (Thompson *et al*., 1994) and the indel positions that were introduced by the alignment software were put into the DNA sequences according to where they were found in the protein alignment. Any amino acid position that could be aligned in more than one way was removed from the alignment. Also, some sequences were removed from the final alignment as a result of alignment difficulty. The result was a 573 bp alignment of protein-coding DNA sequences, where every position in the alignment was aligned with a reasonable degree of confidence.

Analyses of signatures of adaptive evolution were carried out using four different methods. The first was the pair-wise method comparing the *d*N and dS values (Li, 1993). The second method involved the modification of this approach as described by Messier and Stewart (1997), where all the hypothetical ancestral sequences are reconstructed at each internal node of the phylogeny and *d*N and *d*S values are calculated between each node and its descendants. The third method was the relative rate ratio method (Creevey and McInerney, 2002) and the fourth method used the ML approach (Goldman and Yang, 1994; Yang *et al*., 2000b; Yang and Swanson, 2002). The first three methods are implemented in the program CRANN (Creevey and McInerney, 2003) and the ML calculations were carried out using the program package PAML (Yang, 1997).

3.2. Results

3.2.1. Pair-wise Distance Method

A pair-wise distance analysis was carried out for all possible pairs of sequences. In each case the averaged *d*N and the *d*S values were calculated. For just 205 of the total of 1275 comparisons, the *d*N value was greater than the *d*S value. Among these 205 cases, 149 (73%) involved just three sequences: PFD0070, PFD0125 and PFD1220. An examination of the phylogenetic position of these three sequences (Figure 1) indicates that these sequences are closely related to one another. This result provides strong evidence that the evolution of these sequences has been influenced by positive selection.

3.2.2. Messier and Stewart Method

Using Messier and Stewart's method (Messier and Stewart, 1997), two analyses were carried out. Two phylogenetic trees were constructed, one

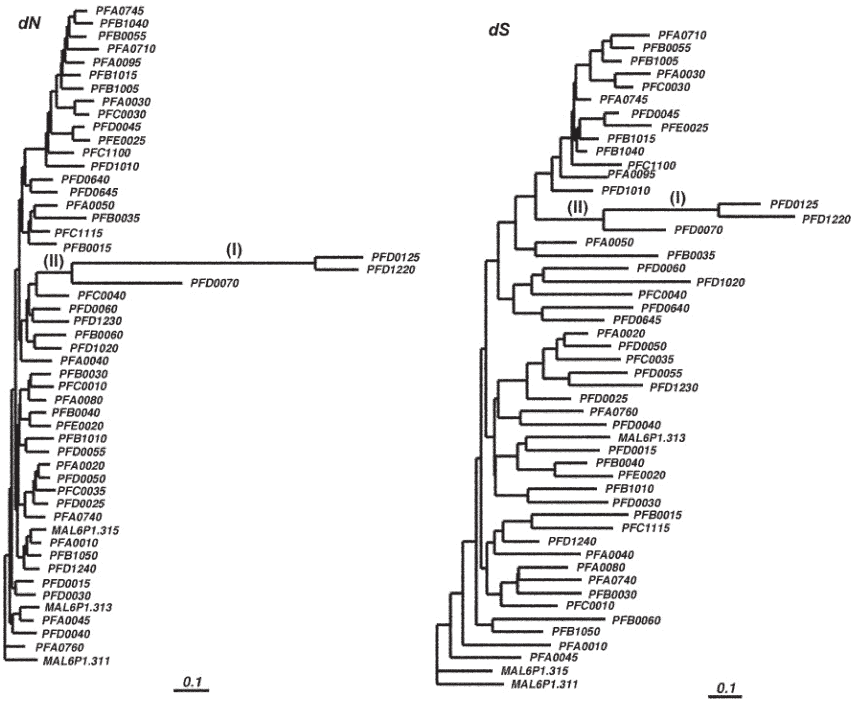

Figure 1 Phylogenetic trees of the *rifin* protein family from *P. falciparum*. The tree on the left has been inferred by the neighbor-joining method based upon dN distances calculated using the Li (1993) method. The tree on the right is based upon dS distances calculated using the same method. Scale bars represent 0.1 substitutions per site.

using the neighbor-joining method based on distances derived from synonymous sites (dS-distances) and one using the neighbor-joining method based on distances derived from nonsynonymous sites (dN-distances) calculated as described above. Both trees are shown in Figure 1. Although the trees differ slightly in their topology, there is broad agreement among the groupings on both trees. Of interest are the branches leading to the three sequences mentioned previously (PFD0070, PFD0125 and PFD1220). This clade appears to be rapidly evolving in both trees, however the branches in this clade are much longer on the tree based on dN distances, relative to the branch lengths in the rest of the tree. For each internal node on the tree, hypothetical ancestral sequences were inferred using parsimonious character reconstruction. Using the dN tree, the estimated $d\mathrm{N}:d\mathrm{S}$ ratio comparing

PFD0070 with the hypothetical common ancestor of the clade was 6.5. The estimated $d\mathrm{N}:d\mathrm{S}$ ratio for the comparison of the common ancestor of PFD1220 and PFD0125 and the common ancestor of the entire clade was 4.3. The $d\mathrm{N}:d\mathrm{S}$ ratio of the common ancestor of PFD0125 and PFD1220 and the extant sequence of PFD0125 was 2.37, while the $d\mathrm{N}:d\mathrm{S}$ ratio for the common ancestor of these two sequences and PFD1220 was less than unity.

3.2.3. Relative Rate Ratio Test (RRRT)

The method proposed by Creevey and McInerney (2002) represents a different kind of approach to the problem of identifying adaptive evolution. In this approach, there is no attempt to estimate $d\mathrm{N}$ and $d\mathrm{S}$ ratios, rather the attempt is to detect clade-specific deviations from neutrality. An entire clade is chosen and the mutational process from the internal branch that describes the clade to the tips of the clade are evaluated. Using this method, and choosing a variety of outgroups, the clade that describes the sequences PFD1220, and PFD0125 was again consistently selected as one that has potentially undergone a period of adaptive evolution. Interestingly the clade that included PFD0070 was not deemed significantly different from neutrality. This indicates that the most probable location of an adaptive evolutionary event is on the internal branch separating PFD1220 and PFD0125 from the rest, not on the internal branch that separates the clade that also includes PFD0070.

3.2.4. Likelihood Ratio Test (LRT) of positive Selection

A variety of analyses was carried out on the alignment using different models of sequence evolution. The results of these analyses are summarized in Table 1. There were four different categories of models employed in these analyses. The first kind of model, designated M0 in Table 1, is the model of Goldman and Yang (1994). In this model a single $d\mathrm{N}:d\mathrm{S}$ ratio (designated ω) is assumed for all positions in the alignment across all lineages. Using this model, the ω value was estimated to be 0.1859, indicating strong purifying selection. However, this model is somewhat unrealistic and restrictive and represents an average across all positions in the alignment and all lineages, so a number of other models were also employed.

The next model was that of Yang (1998). This model allows the ω value to vary across the tree. In this model, we allow a single ω value for background lineages and an ω value for 'foreground' lineages. These ω values are iteratively optimized during the procedure. From our initial analyses

Table 1 Results of ML analysis of the rifin data using a variety of ML models. p_0, p_1, p_2 and p_3 refer to the proportion of sites in categories 0, 1, 2 and 3 respectively. ω refers to the $dN:dS$ ratio in these categories of sites. Descriptions of the various models are to be found in the text.

	p	ln L	Estimates of parameters
M0: one-ratio	1	−14735.843499	$\omega = 0.1859$
Branch-specific		**(PFD0125, PFD1220)**	
Two Ratios	2	−14734.339974	$\omega_0 = 0.1864$ $\omega_1 = 0.0097$
Branch-specific		**(PFD0070, (PFD0125, PFD1220))**	
	2	−14731.799842	$\omega_0 = 0.1868$ $\omega_1 = 0.0014$
Site-specific			
M1–Neutral	1	−15278.599347	$p_0 = 0.05446$ $p_1 = 0.94554$
M2–Selection	3	−14280.751708	$p_0 = 0.04307$ $p_1 = 0.28258$ $p_2 = 0.67435$ $\omega_2 = 0.13284$
M3 – $k = 2$	3	−14219.571671	$p_0 = 0.57091$ $\omega_0 = 0.06189$ $\omega_1 = 0.45720$
M3 – $k = 3$	5	−14134.152650	$p_0 = 0.46970$ $p_1 = 0.37161$ $p_2 = 0.15869$ $\omega_0 = 0.04819$ $\omega_1 = 0.27168$ $\omega_2 = 0.80995$
M7 – beta	2	−14108.801679	$p = 0.56930$ $q = 1.55267$
M8 – beta&ω	4	−14095.766769	$p_0 = 0.90962$ $(p = 0.80147$ $q = 3.53496)$ $\omega = 1.02483$
Branch-site		**(PFD0125, PFD1220)**	
Model A	3	−15217.403731	$P_0 = 0.03549$ $p_1 = 0.43042$ $p_2 = 0.04068$ $p_3 = 0.49341$ $\omega_2 = 128.05992$
Model B	5	−14208.574035	$P_0 = 0.21796$ $p_1 = 0.16293$ $p_2 = 0.35428$ $p_3 = 0.26484$ $\omega_0 = 0.06207$ $\omega_1 = 0.45667$ $\omega_2 = 32.01223$
Branch-site		**(PFD0070, (PFD0125, PFD1220))**	
Model A	3	−15247.783057	$p_0 = 0.04589$ $p_1 = 0.65750$ $p_2 = 0.01935$ $p_3 = 0.27725$ $\omega_2 = 561.08899$
Model B	5	−14209.806432	$P_0 = 0.47544$ $p_1 = 0.34650$ $p_2 = 0.10300$ $p_3 = 0.07507$ $\omega_0 = 0.06123$ $\omega_1 = 0.46200$ $\omega_2 = 519.22796$

described above, we found that three sequences were potentially interesting, so we have chosen to perform different analyses of clades that contain these sequences. The first branch (branch I in Table 1 and Figure 1) was labelled as the foreground lineage and surprisingly, while the background lineages were estimated to have an ω value similar to M0 (0.1864), the foreground lineage was estimated to have an ω value of 0.0097, with a consequent, though not significant, increase in the log-likelihood. A similar situation was observed when the branch leading to the larger clade (branch II in Table 1 and Figure 1) was labeled.

The next series of models that were employed were those models that allow the ω value to vary across the sequence. These models allow for heterogeneity in selective pressures at different amino acid sites, however they do not account for rate variation across lineages. Model M1 is the least realistic of the models. It includes two classes of sites, one class where nonsynonymous substitutions are completely deleterious and removed $\omega = 0$ and one class that are neutral and have an ω value of 1. This particular model had the lowest log-likelihood score of all (-15278.59935). The next model (M2) includes an additional class of site with ω estimated from the data. The likelihood for this model is better than the simpler model M1 and the estimated ω for the additional class of sites is 0.13284, with an estimated proportion of 0.67435 sites in this category. This indicates that there is variation in selective pressure across these sites that is not accounted for by the neutral model.

We then used two models that use an unconstrained discrete distribution to model variability at different sites (M3). In the first of these models, we allow two separate discrete classes of sites ($k = 2$). This model fits the data much better that the one ratio model as evidenced by the increase of approximately 516 log-likelihood units. The M3 model with three site classes is an even better fit to the data by an additional 85 log-likelihood units. The model M7 does not allow for positively selected sites, but allows variation in ω across sites in the interval (0, 1), however it again is a better fit to the data than any model so far. The model that appears to be the best fit is the model M8. This model can be compared with M7 using a χ^2 test with d.f. $= 2$ (i.e. $2\Delta l = [2 \times 13] = 26$, $P < 10^{-4}$). This is much greater than the χ^2 critical value. This model does not identify any amino acid positions that have a high probability of being under positive selection.

The conclusion so far is that models that allow variation in selective pressure across the alignment are preferable to those models that do not allow differences in ω at different sites.

The next category of models we have examined are those mixed models that allow ω to vary among sites and also allow ω to vary in different parts of the tree. These are designated 'Model A' and 'Model B' in Table 1. Given

that we may select different lineages as the 'foreground' and 'background' lineages, we have decided to examine the same two internal branches that were examined in the previous branch-specific models.

Model A is a special case of the site-specific neutral model M1 and can be compared with this model using a χ^2 test with d.f. $= 2$. Model B is a special case of the discrete model M3 with $k = 2$ and can be compared with this model using a χ^2 test with d.f. $= 2$. In all cases, the branch-site models are significantly better than the respective null models with all p-values $< 10^{-4}$. This indicates that models that account for differences along these two lineages are a significantly better fit to the data than models that do not account for this variation. The best fitting branch-site model is model B. This model does not fix any category of site to any particular value. We note, however, that there are significant differences in the outcomes of labelling the different branches.

When the branch leading to the larger clade is labelled as the foreground branch, approximately 17% of sites are predicted to be under strong positive selection ($p_2 + p_3$). When the branch leading to the smaller clade is labelled, a much larger proportion of sites are predicted to be under positive selection. In fact, 61% of the sites are predicted to be under positive selection in this lineage. This result may help to explain the result from the RRRT and the unusual branch lengths for branch I on the dN tree in Figure 1.

3.2.5. Bayesian Inference of Codons under Positive Selection

The branch-site models in our analyses have proposed that a large proportion of sites are under the influence of positive selection. The sites with a high probability of being in the classes of site that are under positive selection are outlined in Table 2. Clearly, when branch I is labelled, a much higher number of sites have a significant probability of being in the class of site that is undergoing positive selection. A total of 28 sites have a probability of greater than 0.95 of being in the class of site that is undergoing positive selection in the internal branch labelled (I) in Figure 1. There are only two sites with this high a probability of being under positive selection on the branch labelled (II) in Figure 1.

4. PROSPECTS

Episodic adaptation to new ecological niches has been reported in a wide variety of proteins, ranging from primate stomach proteins (Messier and

Table 2 A list of the amino acid positions that have a probability greater than 0.95 of having been under positive selection along either of the two branches labelled in Figure 1. Branch I refers to the branch labelled (I) in Figure 1, branch II refers to the branch labelled (II) in Figure 1.

Branch	Branch II
7 N 0.9800*	119 S 0.9917**
10 L 0.9978**	174 R 0.9925**
19 L 0.9900**	
24 C 0.9635*	
31 P 0.9940**	
54 R 0.9784*	
76 D 0.9588*	
82 I 0.9818*	
86 I 0.9927**	
92 E 0.9783*	
93 K 0.9968**	
101 T 0.9960**	
102 L 0.9990**	
113 T 0.9977**	
114 C 0.9969**	
116 C 0.9705*	
118 K 0.9833*	
121 A 0.9978**	
131 C 0.9898*	
136 G 0.9958**	
153 S 0.9960**	
168 Y 0.9967**	
170 I 0.9938**	
172 R 0.9616*	
179 V 0.9872*	
181 K 0.9992**	
188 L b0.9815*	
191 E 0.9849*	

Stewart, 1997) to invertebrate sperm proteins (Swanson and Vacquier, 1995; Galindo *et al*., 2003), and bacterial membrane proteins (Smith *et al*., 1995). In all cases, these adaptive events have been associated with lifestyle changes in the organism in which these changes are found. In this chapter, we have analysed a set of 51 paralagous genes from *P. falciparum*. In three of these genes, we have found evidence that positive selection has

occurred. We have also used LRTs in order to evaluate different models of sequence evolution. The preferred models directly point to adaptive evolution in the clade that defines these sequences. In addition, this finding is supported by pair-wise analyses of the sequences and a relative rate ratio test of deviation from neutrality. We have also identified amino acids within these proteins that are likely to be under positive selection for change.

It is interesting to note that only three proteins, or 6% of the members of this family, appear to be under positive selection for change. Overall, the proteins are not too dissimilar, with typical pair-wise amino acid distances being less than 0.3 substitutions per site. The rifin proteins generally elicit an antibody response from malaria-infected individuals and presumably, because they are recognized by the immune system, they are exposed to strong selective pressures. The reason for positive selection in just these three members of the family may therefore be their location on the cell surface, their level of expression, or perhaps their slightly different function. Based on sequence analysis alone, it is not possible to say with certainty why there are differences in selective pressure among different members of this family, however there is some strong evidence that different members interact differently with the human immune system.

In a study of individuals living in an area where malaria is endemic, Abdel-Latif *et al.* (2002) found that one particular rifin protein caused a greater induction of anti-rifin antibodies than other rifin proteins in the same study. This could suggest that this gene belongs to a subset of stably expressed dominant and commonly recognized *rif* proteins, which are present on the infected red blood cell surface. If this were the case, members of this subset would be continually under selective pressure for change, unlike the less commonly recognized proteins. The results described here of variable selective pressures in different members of the rifin gene family support this theory. This study has identified three rifin proteins that appear to be under positive selection for change. This contrasts with the rest of the members of the family. Positive selection will manifest itself only if the selective advantage offered by the mutation is sufficient to overcome random genetic drift. If a protein is only transiently expressed or if it does not elicit a strong antibody response, then it will not be subject to the same selective pressure as is constitutively expressed and/or is capable of inducing a strong adaptive immune response.

The analyses described in this paper have succeeded in finding a lineage within a multigene family where positive selection appears to characterize its evolutionary history. The rate of retention of replacement substitutions in three genes of the RIFIN family has been greater than the rate of retention of silent substitutions. In other words, those lineages of *P. falciparum* in which these new mutations occurred had a selective advantage as a result.

Those lineages in which these substitutions did not occur were less fit and are not part of the extant genome of *P. falciparum*.

Although it is dangerous to conclude that evidence of adaptive evolution is indicative of lifestyle change, analyses of adaptive evolution may provide testable hypotheses. Subsequently, it may be possible to investigate further (using biochemical or other means) amino acids that are identified as having been under positive selection.

Frank (2002) recently reviewed the advantages of measuring selection in understanding further the immunology and evolution of infectious diseases. In the absence of structural data, detecting positive selection through gene sequence analysis has enabled the prediction of 'which site may be structurally exposed and can change and which sites are either not exposed or functionally constrained' (Frank, 2002). This has led to the recognition of epitopes, the sites within macromolecules to which a specific antibody binds, as well as other positively selected amino acid sites that may or may not constitute unidentified epitopes (Endo *et al*., 1996; Yang and Bielawski, 2000). Sites under strong positive selection may represent candidate vaccination targets (Suzuki and Gojobori, 2001), although there is no means of knowing at present, whether such sites will continue to be primary targets of selection and thus to what extent they will either change or remain important in the host–parasite interaction. Identifying genes and gene regions under selection brings the effect of the environment on the genotype into focus. When selection within a parasite's genome can be demonstrated to be in response to the host, or indeed when host genes can be demonstrated to be under selection in response to a parasite, we can narrow our focus on the evolutionary genetic basis of the host–parasite interaction. For vaccine development there may come a time when routine genome analysis involves the assessment of whether genes appear to be under positive selection even before the function of the genes is fully understood. Phylogenetically based methods of detecting selection within and between lineages clearly have great potential in elucidating gene and protein evolution, and host–parasite interactions. Thus, once again, an accurate phylogeny is pivotal, whether founded on proteins, protein-coding genes, or species.

ACKNOWLEDGEMENTS

We thank Gayle Philip and Peter G. Foster for their help with this manuscript. J.O.M. and C.J.C. acknowledge the financial support of the Irish Research Council for Science, Engineering and Technology, Enterprise

Ireland, The Health Research Board (H.R.B) and H.E.A. PRTLI Cycle II and III. D.T.J.L. was funded by a Wellcome Trust Fellowship (043965). We thank two anonymous reviewers for their invaluable positive contribution to this manuscript.

REFERENCES

Abdel-Latif, M.S., Khattab, A., Lindenthal, C., Kremsner, P.G. and Klinkert, M.-Q. (2002). Recognition of variant rifin antigens by human antibodies induced during natural *Plasmodium falciparum* infections. *Infection and Immunity* **70**, 7013–7021.

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research* **25**, 3389–3402.

Bielawski, J.P. and Yang, Z. (2001). Positive and negative selection in the DAZ gene family. *Molecular Biology and Evolution* **18**, 523–529.

Creevey, C. and McInerney, J.O. (2002). An algorithm for detecting directional and non-directional positive selection, neutrality and negative selection in protein coding DNA sequences. *Gene* **300**, 43–51.

Creevey, C. and McInerney, J.O. (2003). CRANN: a program for detecting positive selection in protein-coding genes. *Bioinformatics*, in press.

Creighton, T.E. and Darby, N.J. (1989). Functional evolutionary divergence of proteolytic enzymes and their inhibitors. *Trends in Biochemical Sciences* **14**, 319–324.

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

Felsenstein, J. (1981). Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of Molecular Evolution* **17**, 368–376.

Felsenstein, J. (1996). Inferring phylogenies from protein sequences by parsimony, distance and likelihood methods. In: *Computer Methods for Macromolecular Sequence Analysis* (R.F. Doolittle, ed.), *Methods in Enzymology*, Vol. 266, pp. 418–427. Orlando: Academic Press.

Fernandez, V., Hommel, M., Chen, Q., Hagblom, P. and Wahlgren, M. (1999). Small, clonally variant antigens expressed on the surface of the *Plasmodium falciparum*-infected erythrocyte are encoded by the *rif* gene family and are the target of human immune responses. *Journal of Experimental Medicine* **190**, 1393–1403.

Forsyth, K.P., Philip, G., Smith, T., Kum, E., Southwell, B. and Brown, G.V. (1989). Diversity of antigens expressed on the surface of erythrocytes infected with mature *Plasmodium falciparum* parasites in Papua New Guinea. *American Journal of Tropical Medicine and Hygiene* **41**, 259–265.

Frank, S.A. (2002). *Immunology and Evolution of Infectious Diseases*. Princeton: Princeton University Press.

Galindo, B.E., Vacquier, V.D. and Swanson, W.J. (2003). Positive selection in the egg receptor for abalone sperm lysin. *Proceedings of the National Academy of Sciences of the USA* **100**, 4639–4643.

Gardner, M.J. and 26 others (1998). Chromosome 2 sequence of the human malaria parasite *Plasmodium falciparum*. *Science* **283**, 1126–1132.

Gaut, B.S. and Clegg, M.T. (1991). Molecular evolution of alcohol dehydrogenase 1 in members of the grass family. *Proceedings of the National Academy of Sciences of the USA* **88**, 2060–2064.

Gaut, B.S. and Clegg, M.T. (1993). Molecular evolution of the adh1 locus in the genus *Zea*. *Proceedings of the National Academy of Sciences of the USA* **90**, 5095–5099.

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

Good, M.F. and Doolan, D.L. (1999). Immune effector mechanisms in malaria. *Current Opinions in Immunology* **11**, 412–419.

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

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

Hughes, M.K. and Hughes, A.L. (1995). Natural selection on *Plasmodium* surface proteins. *Molecular and Biochemical Parasitology* **71**, 99–113.

Ina, Y. (1995). New methods for estimating the numbers of synonymous and nonsynonymous substitutions. *Journal of Molecular Evolution* **40**, 190–226.

Kimura, M. (1968). Evolutionary rate at the molecular level. *Nature* **217**, 624–626.

Kyes, S.A., Rowe, J.A., Kriek, N. and Newbold, C.I. (1999). Rifins: a second family of clonally variant proteins expressed on the surface of red cells infected with *Plasmodium falciparum*. *Proceedings of the National Academy of Sciences of the USA* **96**, 9333–9338.

Kyes, S., Pinches, R. and Newbold, C. (2000). A simple RNA analysis method shows *var* and *rif* multigene family expression patterns in *Plasmodium falciparum*. *Molecular and Biochemical Parasitology* **105**, 311–315.

Li, W.-H. (1993). Unbiased estimation of the rates of synonymous and nonsynonymous substitution. *Journal of Molecular Evolution* **36**, 96–99.

Li, W.-H., Wu, C.-I. and Luo, C.-C. (1985). A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution and considering the relative likelihood of nucleotide and codon changes. *Molecular Biology and Evolution* **2**, 150–174.

Maynard-Smith, J. (1970). Population size, polymorphism, and the rate of non-Darwinian evolution. *American Naturalist* **104**, 231–236.

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

McInerney, J.O. (1998). Replicational and transcriptional selection on codon usage in *Borrelia burgdorferi*. *Proceedings of the National Academy of Sciences of the USA* **95**, 10698–10703.

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

Nei, M. and Gojobori, T. (1986). Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Molecular Biology and Evolution* **3**, 418–426.

Nielsen, R. and Yang, Z. (1998). Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. *Genetics* **134**, 1271–1276.

Smith, N.H., Smith, J.M. and Spratt, B.G. (1995). Sequence evolution of the *porb* gene of *Neisseria gonorrhoeae* and *Neisseria meningitidis* – evidence of positive Darwinian selection. *Molecular Biology and Evolution* **12**, 363–370.

Suzuki, Y. and Gojobori, T. (2001). Positively selected amino acid sites in the entire coding region of hepatitis C virus subtype 1b. *Gene* **276**, 83–87.

Swanson, W.J. and Vacquier, V.D. (1995). Extraordinary divergence and positive Darwinian selection in a fusagenic protein coating the acrosomal process of abalone spermatozoa. *Proceedings of the National Academy of Sciences of the USA* **92**, 4957–4961.

Thompson, J.D., Higgins, D.G. and Gibson, T.J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research* **22**, 4673–4680.

Weber, J.L. (1988). Interspersed repetitive DNA from *Plasmodium falciparum*. *Molecular and Biochemical Parasitology* **29**, 117–124.

Winzeler, E.A., Liang, H., Shoemaker, D.D. and Davis, R.W. (2000). Functional analysis of the yeast genome by precise deletion and parallel phenotypic characterization. *Novartis Foundation Symposium* **229**, 105–109; discussion pp. 109–111.

World Health Organization (1997). World malaria situation in 1994. Part I. Population at risk. *Weekly Epidemiological Record* **72**, 269–274.

World Health Organization (2002). Economic costs of malaria. *Roll Back Malaria Information*, sheet 10 of 11.

Yang, Z. (1997). PAML: a program package for phylogenetic analysis by maximum likelihood. *Computer Applications in the Biosciences* **13**, 555–556.

Yang, Z. (1998). Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. *Molecular Biology and Evolution* **15**, 568–573.

Yang, Z. (2000). Maximum likelihood estimation on large phylogenies and analysis of adaptive evolution in human influenza virus A. *Journal of Molecular Evolution* **51**, 423–432.

Yang, Z. and Bielawski, J.P. (2000). Statistical methods for detecting molecular adaptation. *Trends in Ecology and Evolution* **15**, 496–503.

Yang, Z. and Nielsen, R. (2002). Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. *Molecular Biology and Evolution* **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. *Molecular Biology and Evolution* **19**, 49–57.

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

Yang, Z., Swanson, W.J. and Vacquier, V.D. (2000a). Maximum-likelihood analysis of molecular adaptation in abalone sperm lysin reveals variable selective pressures among lineages and sites. *Molecular Biology and Evolution* **17**, 1446–1455.

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

Zanotto, P.M., 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.