Chapter 4

Evolutionary Model of Immune Selection

For a parasite such as *Neisseria meningitidis*, the key to long-term persistence is the successful and ongoing colonisation of a host. Despite its notorious pathogenicity, *N. meningitidis* normally resides as a commensal of the nasopharynx, but that is not to say that *N. meningitidis* is an onlooker in the co-evolutionary arms race between the host immune system and microbial intruders. Antigenic variation is a distinguishing feature of meningococcal populations, indicating that the observed genetic diversity at these loci is caused by strong selective pressures. Indeed, the patterns of genetic variation in samples of antigen gene sequences can be used to locate individual sites that interact directly with the host immune system.

Current methods that attempt to identify sites that interact with the immune system are based on reconstructing the phylogenetic tree of the gene sequences. In a highly recombining organism such as *N. meningitidis*, phylogenetic methods are not appropriate because there may be multiple trees along the sequence. In the presence of high levels of recombination phylogenetic methods that attempt to detect positive selection can have a false positive rate of up to 90% (Anisimova *et al.* 2003; Shriner *et al.* 2003). In this chapter I will begin by discussing the background to the dN/dS ratio (section 4.1.1), and the current phylogenetic methods for detecting immune selection (section 4.1.2). In section 4.2 I present a new population genetics model of immune selection in the presence of recombination, based on an approximation to the coalescent (Li and Stephens 2003). I also describe a model for variation in selection

pressure and the recombination rate within a gene, which is novel in the context of detecting selection. In section 4.3 I describe how to perform Bayesian inference on the selection and recombination parameters under the new model, using reversible-jump Markov chain Monte Carlo (MCMC). Then in section 4.4, I use a simulation study to investigate the properties of the inference method under two scenarios and demonstrate that the new method has the power to detect variability in selection pressure and recombination rate, and does not suffer from a high false positive rate. In Chapter 5 I apply the new method to the *porB* locus of *N. meningitidis* which encodes the PorB outer membrane protein. I use prior sensitivity analysis and model criticism techniques to verify the inferences, and compare the results to those obtained with phylogenetic methods.

4.1 The dN/dS ratio

4.1.1 Models that incorporate the dN/dS ratio

As an indicator of the action of natural selection in gene sequences the ratio of non-synonymous to synonymous substitutions (dN/dS) is a versatile and widely-used method of summarising patterns of genetic diversity. When comparing a pair of nucleotide sequences, synonymous substitutions refer to those codons that differ in their nucleotide sequence but not in the amino acid encoded. Non-synonymous substitutions refer to those codons that differ both in nucleotide sequence and amino acid encoded. In Figure 1 there are nine possible single nucleotide mutations of the triplet `CTT`, two of which are synonymous because leucine is still encoded, the other seven of which are non-synonymous.

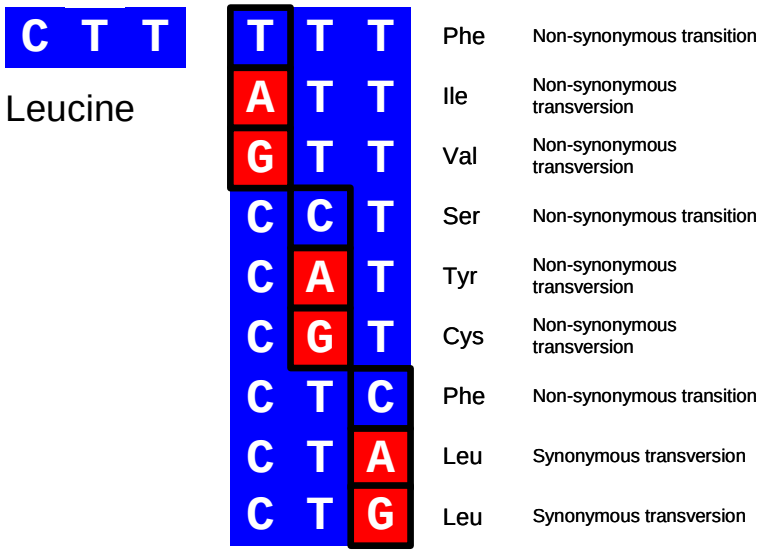

Figure 1 Synonymous and non-synonymous single nucleotide mutations from CTT.

In a strictly neutral model in which single nucleotide mutations occur at a uniform rate, non-synonymous mutations would occur more frequently than synonymous mutations, because there are more potential non-synonymous changes. The dN/dS ratio measures the relative rate at which non-synonymous and synonymous changes occur, adjusting for the fact that there are more potential non-synonymous changes. In a strictly neutral model of evolution, the dN/dS ratio equals one. The dN/dS ratio is an indicator of natural selection, because deviations from a ratio of one suggest that nucleotide changes that alter the amino acid sequence are more or less frequently observed than those that do not.

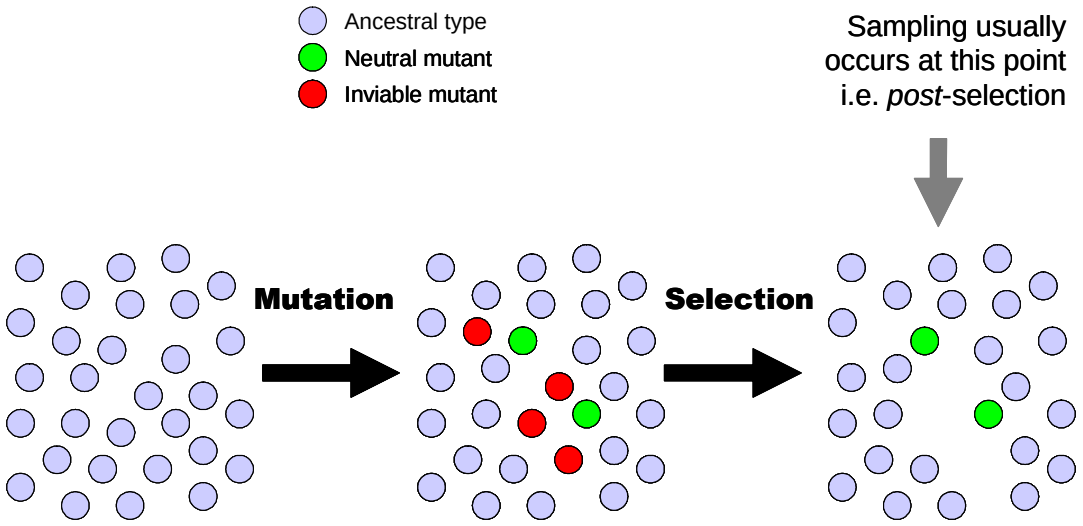

Figure 2 In samples of gene sequences the effects of mutation and selection on patterns of genetic diversity are confounded. For example, non-synonymous polymorphism might be under-represented because of purifying selection.

4.1.1.1 Purifying selection and dN/dS

Figure 2 illustrates how the observed patterns of synonymous and non-synonymous polymorphism represent a confounding between the evolutionary processes of mutation and natural selection. For example, it is generally assumed in studies of adaptation that organisms are optimally adapted to their environment (Dawkins 1982). This is a reasonable assumption because over long periods of time natural selection favours variants that have a selective advantage. If a gene is adapted to its environment, even if it is not optimally adapted, then there will be a great many more worse alternative sequences than better alternative sequences. So, random mutation will tend to produce less-well adapted sequences, not better adapted sequences. As a result of natural selection, those sequences that have reduced survival or reproductive success will be under-represented in a sample taken from the population. None of this applies to synonymous changes, of course, which do not alter the amino acid sequence of the gene product. As a result it is reasonable to expect that purifying, or

negative, selection will cause non-synonymous variants to be under-represented relative to synonymous variants, and the dN/dS ratio will be less than one in a functional gene. This is known as functional constraint.

The fact that mutation and natural selection are confounded in genetic samples serves as the basis for a class of evolutionary models of selection. Models of selection that describe the movement of alleles through the population (e.g. Fisher 1930) are not easily amenable to inference because for each site and each allele the selective advantage conferred by that allele (the selection coefficient), the time since the allele arose, and the way in which selection coefficients interact across sites, all need to be specified, resulting in a great many parameters. Such models exist, usually they make assumptions to reduce the number of parameters, but the inference methods are computationally prohibitive even when recombination is not modelled (e.g. Coop and Griffiths 2004). Evolutionary models that deliberately confound mutation and natural selection (Goldman and Yang 1994; Nielsen and Yang 1998; Sainudiin *et al.* 2005) use a single selection parameter for each site, the dN/dS ratio. In these models natural selection is treated as a form of mutational bias, so that if the dN/dS ratio is less than one then non-synonymous mutations simply occur at a lower rate.

In the codon model of Nielsen and Yang (1998), hereafter NY98, the mutation rate from codon i to j ($i \neq j$), which I will measure in units of PN_e generations (where P is the ploidy and N_e the effective population size) is

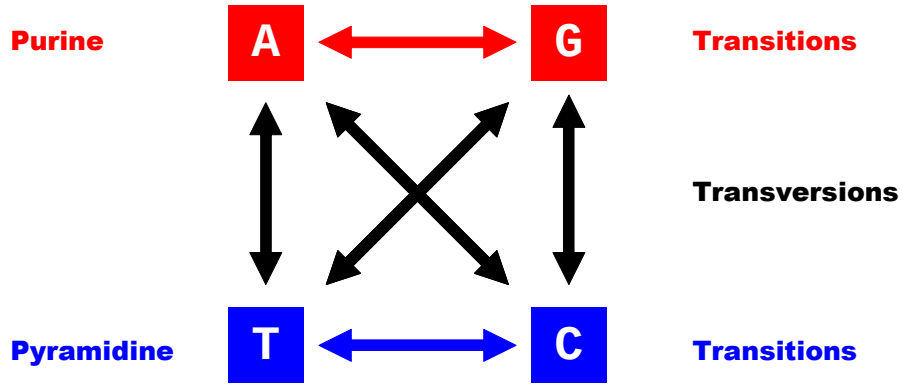

Figure 3 There are two classes of nucleotides, purines (adenosine and guanine) and pyrimidines (thymine, cytosine and uracil). Single nucleotide mutations that do not change the nucleotide class are called transitions, and those that do are called transversions. For any nucleotide there are two possible transversions and one transition. Despite this, transitions are observed more commonly than transversions, so the transition:transversion ratio κ is usually greater than two.

$$q_{ij} = \pi_j \mu \begin{cases} 1 & \text{if i and j differ by a synonymous transversion} \\ \kappa & \text{if i and j differ by a synonymous transition} \\ \omega & \text{if i and j differ by a nonsynonymous transversion,} \\ \kappa\omega & \text{if i and j differ by a nonsynonymous transition} \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

and $q_{ii} = -\sum_{j \neq i} q_{ij}$, where the frequency of codon j is π_j, κ is the relative rate of transitions to transversions (defined in Figure 3), and ω is the dN/dS ratio. If there were equal codon usage (i.e. $\pi_j = 1/61$ because only the 61 non-stop codons are allowed in NY98) the total rate of synonymous mutation (per PN_e generations) would be approximately,

$$\frac{\theta_s}{2} \approx \frac{(6+5\kappa)\mu}{310}. \quad (2)$$

4.1.1.2 Positive selection and dN/dS

When organisms are already well-adapted to their environment natural selection will purge the population of less-fit variant genes so non-synonymous polymorphism is under-represented relative to synonymous polymorphism and $dN/dS < 1$. The converse scenario, in which non-synonymous polymorphism is over-represented relative to synonymous polymorphism and $dN/dS > 1$ needs careful interpretation. An excess of non-synonymous polymorphism implies that there is a selective advantage to novelty in the amino acid sequence. It might be envisaged that recurrent, adaptive change in a gene will manifest itself as an over-representation of non-synonymous relative to synonymous change because positive selection will drive the adaptive variants to high frequency. Such a model has been used to detect natural selection between species, because it is assumed that multiple adaptive changes are important during speciation (McDonald and Kreitman 1991; Shpaer and Mullins 1993; Long and Langley 1993).

However, some controversy surrounds the generality with which adaptation leads to an excess of non-synonymous polymorphism. When positive dN/dS is observed, it is likely that multiple compensatory, or complementary, changes at several sites in the gene have occurred as a result of adaptation. So an excess of non-synonymous relative to synonymous polymorphism is a clear signal of adaptive change, or positive selection. But a single adaptive substitution at a particular codon is not sufficient to generate a positive dN/dS ratio across a whole gene if much of the gene is functionally constrained. So the dN/dS ratio will under-report the extent of adaptive change for any model in which episodic environmental change causes a transformation from one optimal state to a new optimum.

What an excess of non-synonymous to synonymous polymorphism is truly indicative of is selection for variation in the polypeptide sequence, not change from one conserved state to another. That makes the dN/dS ratio a particularly useful tool for studying the interaction between antigen genes and the immune system. Immunological memory against particular antigens exerts a strong selective pressure for antigenic novelty in the parasite population. This is known as diversifying selection. The antigenic properties of an outer membrane protein such as PorB may be determined by a small number of amino acids that might or might not be contiguous in the codon sequence. The dN/dS ratio can in principle be harnessed to estimate the magnitude of the selection pressure exerted by the immune system on different genes, investigate the evolutionary trade-off between protein functionality and immune evasion in the parasite, and locate the genetic determinants of antigenicity at a locus. The latter might be informative for vaccine development.

4.1.2 Inferring immune selection using dN/dS

Nielsen and Yang (1998) proposed a maximum likelihood phylogenetic approach to estimating the dN/dS ratio that employs a codon-based mutation model (Equation 1), and treats the dN/dS ratio as an unknown parameter ω. This method has subsequently been expanded (Yang *et al*. 2000; Yang and Swanson 2002; Swanson *et al*. 2003), adapted into a Bayesian setting (Huelsenbeck and Dyer 2004), and approximated for the purposes of computational efficiency (Massingham and Goldman 2005). Simulation studies have shown that phylogenetic likelihood-based methods can be substantially more powerful than alternative non-likelihood-based approaches

(Anisimova *et al*. 2001; Anisimova *et al*. 2002; Wong *et al*. 2004; Kosakovsky Pond and Frost 2005).

Estimating the selection parameter ω using these methods has become widespread (e.g. Bishop *et al*. 2000; Ford 2001; Mondragon-Palomino *et al*. 2002; Filip and Mundy 2004) and has been applied to many organisms. Analysis of pathogens such as viruses (Twiddy *et al*. 2002; Moury 2004; de Oliveira *et al*. 2004) and bacteria (Peek *et al*. 2001; Urwin *et al*. 2002) is particularly informative, because they typically have high mutation rates and are consequently genetically diverse, which lends greater statistical power to estimation. As discussed, the diversifying selection imposed by the host immune system may be the most appropriate model for which inference based on the dN/dS ratio can be applied. The ability to observe these populations evolving in real-time makes them especially interesting for the study of evolution (Drummond *et al*. 2003a), and suggests that we may be able to make useful epidemiological inference from molecular sequence data.

4.1.2.1 CODEML

The method of Nielsen and Yang (1998) is the most popular method for estimating the dN/dS ratio for nucleotide sequence data, and has been widely applied to samples within parasite populations. Based on the mutation model specified by Equation 1, in its original incarnation a random effects model is used for variation in ω between sites. To make inference feasible, only three classes of sites, occurring in proportions p_0, p_1 and p_2 are allowed. These have dN/dS ratios ω_0, ω_1 and ω_2 respectively, subject to the constraint that $\omega_0 < \omega_1 < \omega_2$. The method has three stages.

1. A tree topology is supplied or estimated using maximum likelihood (ML) from the data using a simple nucleotide mutation model.
2. Conditional on the topology, the branch lengths, κ, p_0, p_1 and $\boldsymbol{\omega}$ are estimated by maximum likelihood.
3. An empirical Bayes (Robbins 1956) approach is used to obtain the posterior probability that a given site is a member of a particular class.

The posterior probability that site *h* belongs to class *k*, so that the selection parameter at site *h*, w_h say, equals ω_k is taken to be

$$\Pr(w_h = \omega_k \mid \mathbf{X}_h) = \frac{p_k f(\mathbf{X}_h \mid w_h = \omega_k)}{\sum_{l=0}^{2} p_k f(\mathbf{X}_h \mid w_h = \omega_l)}, \quad (2)$$

(Nielsen and Yang 1998) where $\mathbf{X}_h$ is the codon alignment at site *h* and $f(\mathbf{X}_h \mid w_h = \omega_k)$ is the likelihood function. Equation 2 hides some of the conditioning however. The likelihood function in Equation 2 is not marginal to, but conditional upon the ML tree topology, branch lengths and κ, which are estimated using the alignment across all sites, $\mathbf{X}$. The posterior probability of belonging to class *k* is also conditional upon the ML estimates of p_0 and p_1.

The method of Nielsen and Yang (1998) is implemented in the program CODEML, part of the PAML package (Yang 1997). CODEML includes a large number of alternative specifications for the variation in ω over the sequence, including an arbitrary number of classes, gamma, beta and truncated normal distributions for the variation in ω across sites (the distributions have to be discretised for computational feasibility) and combinations thereof (Yang *et al.* 2000). Nielsen and Yang (1998) use a likelihood ratio test to compare nested models of variation in ω. For example, a model with three classes where $\omega_0 = 0$, $\omega_1 = 1$ and $\omega_2 > 1$ can be compared to a model

with only two classes where $\omega_0 = 0$ and $\omega_1 = 1$. This constitutes a test for positive selection. Nielsen and Yang (1998) assume that for nested models, the difference in double the log likelihood (the deviance) follows a χ^2 distribution with degrees of freedom equal to the difference in number of parameters. In practice this asymptotic result might not hold (Anisimova *et al.* 2001).

4.1.2.2 MrBayes

Huelsenbeck and Dyer (2004) implement the model of Nielsen and Yang (1998) in a fully Bayesian setting, available in MrBayes 3 (Ronquist and Huelsenbeck 2003). MrBayes uses MCMC to obtain a posterior distribution for all parameters of the model: codon frequencies, tree topology and branch lengths, κ, **p** and **ω**. Not surprisingly, it is considerably more computationally intensive than CODEML.

Huelsenbeck and Dyer (2004) fit a uniform prior on all unrooted tree topologies, and an exponential prior on branch lengths. Symmetric Dirichlet priors are applied to the frequencies **p** of the ω classes, and the codon frequencies. For the transition:transversion ratio κ, a distribution describing the ratio of two i.i.d. (independently and identically distributed) exponential random variables is used:

$$f(\kappa) = \frac{1}{(1+\kappa)^2}, \kappa > 0 .$$

Under their prior, the selection parameters ω_0, ω_1 and ω_2 are treated as ordered draws ($\omega_0 < \omega_1 < \omega_2$) from a distribution describing the ratio of two i.i.d. exponential random variables:

$$f(\omega_0, \omega_1, \omega_2) = \frac{36}{(1+\omega_0+\omega_1+\omega_2)^4} .$$

Because MrBayes is fully Bayesian, uncertainty in the phylogeny, mutation parameters and ω class frequencies is taken into account in the posterior probability that site *h* belongs to class *k*

$$\Pr(w_h = \omega_k \mid \mathbf{X}) = \int f(w_h = \omega_k, \Theta \mid \mathbf{X}) \mathrm{d}\Theta,$$

where Θ represents $\omega_{[-k]}, \mathbf{p}, \kappa$, the phylogenetic tree topology and branch lengths.

4.1.2.3 SLR

Massingham and Goldman (2005) introduced the sitewise likelihood ratio (SLR) method, which is an approximation to the ML method of Nielsen and Yang (1998). SLR is principally concerned with identifying the mode of selection at each site (i.e. dN/dS < 1 or dN/dS > 1).

The problem with estimating a separate dN/dS ratio for every codon in a sequence is that there are too many parameters. Nielsen and Yang (1998) overcame this problem by using a random effects model for the variation in ω which reduces the number of parameters to only a few. In CODEML, maximum likelihood estimates of the parameters are obtained using a high dimensional optimization procedure which is computationally intensive. In contrast, Massingham and Goldman (2005) use an approximation, described below, that allows a different ω to be estimated for each site. The approximation allows there to be a single multidimensional optimisation for the whole sequence (with fewer parameters than in CODEML) and then a one dimensional optimisation for each site.

The method works as follows

1. Assuming a common selection parameter for the whole sequence ω_0, the maximum likelihood tree, including branch lengths and the parameters κ and ω_0 are jointly estimated.
2. For each site *i* an individual selection parameter ω_i is estimated, assuming that the selection parameter for all other sites is ω_0.
3. For each site *i* a likelihood ratio test is performed for the null hypothesis that $\omega_i = 1$ by assuming that the difference the deviance between $\omega_i = \hat{\omega}_i$ and $\omega_i = 1$ is χ^2 distributed with one degree of freedom.

The method is approximate because for each site ω_i is estimated conditional upon all the other parameters, including ω_0. As a result estimating ω_i is a one dimensional problem for each site. The χ^2 distribution used in the likelihood ratio test is an asymptotic result that may not hold, so a parametric bootstrap procedure (Goldman 1993) can also be used to generate the null distribution of the difference in deviances.

4.1.2.4 Problems with current methods

CODEML, MrBayes and SLR all rely on reconstructions of the phylogenetic tree for the sample of genes. These methods have been applied frequently to within-population samples of micro-organisms (Twiddy *et al*. 2002; Moury 2004; de Oliveira *et al*. 2004; Peek *et al*. 2001; Urwin *et al*. 2002). However, the use of phylogenetic techniques is questionable in organisms that are highly recombining, because recombination leads to not one, but multiple evolutionary trees along the sequence. If the recombination rate is of the same order as the mutation rate, as has been found in some organisms (McVean *et al*. 2002; Stumpf and McVean 2003), then there might be a new evolutionary tree for every polymorphic site along the sequence. In such a scenario, which is plausible for many highly-recombining micro-organisms (Awadalla

2003) and eukaryotic genes containing recombination hotspots (McVean *et al*. 2004, Winckler *et al*. 2005), there is little hope to infer any particular evolutionary tree along the sequence. When a single evolutionary tree is estimated for a sample of gene sequences that have undergone recombination, the resulting tree is likely to have longer terminal branches and total branch length, yet a smaller time to the most recent common ancestor, in a way that superficially resembles the star-shaped topology of an exponentially growing population (Schierup and Hein 2000). The effect on detecting diversifying selection is to produce a high rate of false positives (Anisimova *et al*. 2003), as high as 90% (Shriner *et al*. 2003).

4.2 Modelling selection with recombination

4.2.1 Population genetics inference

When changes in the evolutionary tree are separated by only a few polymorphic sites, there is little hope to infer the tree at any particular site along the sequence. The population genetics approach is to treat the evolutionary trees along the sequence, or genealogy, as missing data. Because the likelihood of a set of molecular sequences needs to be evaluated with reference to a particular genealogy (Felsenstein 1981), it is calculated by averaging over the genealogies, weighted by the probability of that genealogy under the missing data model.

$$P(\mathbf{H} \mid \Theta) = \int P(\mathbf{H} \mid \Theta, G) P(G) \mathrm{d}G , \qquad (3)$$

where $P(\mathbf{H} \mid \Theta)$ is the likelihood of the data **H** given the parameters Θ, $P(G)$ is the missing data model for the genealogy and $P(\mathbf{H} \mid \Theta, G)$ is obtained using the pruning algorithm (Felsenstein 1981). There are various ways to model $P(G)$. In the case of

no recombination Huelsenbeck and Dyer (2004) used a model in which all unrooted tree topologies were uniformly likely, and branch lengths had an exponential distribution. When the sequences are from a single population a natural choice would be the coalescent (Kingman 1982a, 1982b; Hudson 1983; Griffiths and Marjoram 1997) which models a neutrally evolving, randomly mating population of constant size, with or without recombination.

However, $P(\mathbf{H} \mid \Theta, G)$ involves summation over the unknown states of internal nodes in the marginal genealogies (the evolutionary tree at a particular site), so the integration in Equation 3 cannot be solved analytically for any genealogical model, including the coalescent. As a result Equation 3 has to be evaluated numerically, which is not a trivial problem. Naïvely,

$$P(\mathbf{H} \mid \Theta) \approx \frac{1}{M} \sum_{i=1}^{M} P(\mathbf{H} \mid \Theta, G^{(i)}), \tag{4}$$

for large *M*, where $G^{(i)}$ is simulated from $P(G)$. Unfortunately, for all but the simplest problems this method is useless because for most trees drawn from $P(G)$, the conditional likelihood $P(\mathbf{H} \mid \Theta, G)$ is negligibly small. Only once in a million draws would the conditional likelihood contribute significantly to the sum (Stephens 2003).

Importance sampling and Markov Chain Monte Carlo are methods that attempt to calculate Equation 4 more efficiently (see Stephens 2003). Both methods have been applied to a variety of contexts in population genetics (e.g. Kuhner *et al.* 1995, 1998, 2000; Griffiths and Marjoram 1996; Beerli and Felsenstein 1999, 2001; Bahlo and Griffiths 2000; Stephens and Donnelly 2000; Fearnhead and Donnelly 2001; Drummond *et al.* 2002; Wilson *et al.* 2003; Coop and Griffiths 2004; De Iorio *et al.*

2005). The methodology is more tractable in the absence of recombination because the state space of the possible genealogies is much smaller. In the presence of recombination, only the simplest models with two parameters (the mutation rate and recombination rate) have been implemented (Fearnhead and Donnelly 2001; Kuhner *et al.* 2000). Even for a small number of sequences these methods are extremely computationally burdensome. In the context of the NY98 mutation model with variation in the selection parameter and recombination rate amongst sites, such an approach is not feasible.

4.2.2 An approximation to the coalescent

Instead I turn to an approximation to the coalescent likelihood in the presence of recombination (Li and Stephens 2003) called the PAC likelihood ("product of approximate conditionals"). Their approach relies on rewriting the likelihood as

$$P(\mathbf{H} \mid \Theta) = P(H_1 \mid \Theta)P(H_2 \mid H_1, \Theta)\cdots P(H_n \mid H_1, H_2, \ldots, H_{n-1}, \Theta) \qquad (4)$$

where $\mathbf{H} = (H_1, H_2, \ldots, H_n)$ is the sample of n gene sequences (haplotypes). Li and Stephens approximate the $(k+1)$th conditional likelihood

$$P(H_{k+1} \mid H_1, H_2, \ldots, H_k, \Theta) \approx \hat{\pi}(H_{k+1} \mid H_1, H_2, \ldots, H_k, \Theta).$$

The approximate conditional likelihood, $\hat{\pi}$, that they use is a hidden Markov model that is designed to incorporate some key properties of the proper likelihood, notably that (i) the $(k+1)$th haplotype is likely to resemble the first k haplotypes but (ii) recombination means that it may be a mosaic of those haplotypes and (iii) mutation means that it may be an imperfect copy. In terms of averaging over possible evolutionary trees, one can think of the hidden Markov model doing so implicitly, but in an approximate way that is highly computationally efficient.

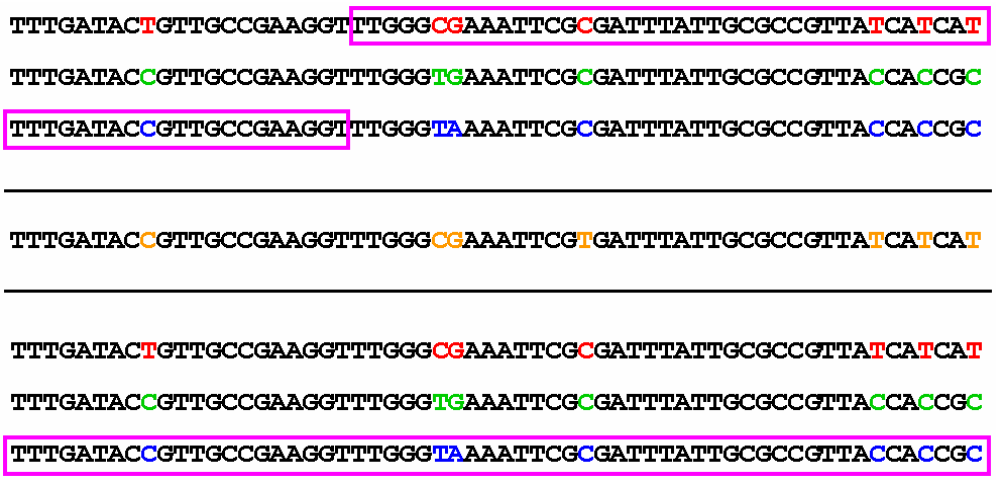

Figure 4 Approximate likelihood of the orange haplotype conditional on the red, green and blue haplotypes. In Li and Stephens' (2003) model, the orange haplotype resembles the others, but recombination means it may be a mosaic and mutation means that it may be an imperfect copy. In the top scenario, the orange haplotype is a mosaic of the red and blue haplotypes, necessitating a C→T mutation. In the bottom scenario, the orange haplotype is a copy of the blue haplotype, necessitating five mutations: T→C, and four C→Ts.

As a result of the approximate nature of the PAC likelihood, the ordering of the n haplotypes can influence the value of the likelihood (were it not for the approximation, the haplotypes would be exchangeable). Therefore, the likelihood is assessed by averaging over multiple orderings of the haplotypes. In the analyses I present throughout this chapter and Chapter 5, I use 10 orderings unless otherwise stated.

4.2.2.1 Sampling formula with recombination

Li and Stephens (2003) use a hidden Markov model (HMM) to approximate the likelihood of the $(k+1)$th haplotype conditional on the first k. Theirs is an approximation to the sampling formula in the sense of Ewens (1972), with the

additional complication of recombination. Li and Stephens think of the $(k+1)$th haplotype as a copy of the first *k* haplotypes. Figure 4 illustrates the idea. At every site, the orange haplotype is a copy of one of the four other haplotypes. This haplotype can be thought of as being closest to the orange haplotype in the evolutionary tree. Parsing the sequence 5′ to 3′, the orange haplotype is a copy of the blue haplotype, so at the first polymorphic site, depending on the mutation rate, it is most likely to share the same nucleotide C. Continuing along the sequence, the orange haplotype can switch between the other four with a given probability. However, if the orange haplotype is a copy of the blue haplotype at site *i*, then it is most likely to continue copying the blue haplotype at site $(i+1)$. This models the way that recombination creates mosaics of contiguous sequences. Between the first and second polymorphic site, the orange haplotype might switch from copying the blue to copying the red haplotype (Figure 4, top). In that case only one mutation need be invoked for the rest of the sequence. However, with some probability the orange continues to copy the blue haplotype (Figure 4, bottom), in which case five more mutation events need to be invoked.

4.2.2.2 Mutation model

In the lexicon of HMMs, the latent variable records which of the first *k* haplotypes the $(k+1)$th is a copy of at a given site. Conditional on the latent variable *x* ($x = 0,1,\ldots,k$), the emission probability models the mutation process, because it specifies the probability of observing state $a = H_{k+1,i}$ in haplotype $(k+1)$ given state $b = H_{x,i}$ in haplotype *x*, at a particular site *i*. Under a coalescent model (Kingman 1981, Hudson 1983), the time (in units of PN_e generations) to the common ancestor of

haplotypes x and $k+1$ is known (R. C. Griffiths, unpublished), and to the order of the approximation is exponentially distributed with rate k. Consider a simple mutation model with two states 0 and 1, and mutation rate $\theta/2$ per PN_e generations. The model is defined by the instantaneous rate matrix

$$\mathbf{Q} = \begin{vmatrix} -\theta/2 & \theta/2 \\ \theta/2 & -\theta/2 \end{vmatrix}. \tag{5}$$

The matrix $\mathbf{P}^{(t)}$ gives the probability $p_{ij}^{(t)}$ of a site being in state j time t after it was in state i.

$$\mathbf{P}^{(t)} = e^{t\mathbf{Q}} \tag{6}$$

(see Grimmett and Stirzaker 2001), which can be solved analytically for this model to give

$$p_{ij}^{(t)} = \begin{cases} \frac{1}{2} + \frac{1}{2}\exp\{-\theta t\} & \text{for } i = j \\ \frac{1}{2} - \frac{1}{2}\exp\{-\theta t\} & \text{for } i \neq j \end{cases}.$$

The probability of observing an (unordered) pair of states (a,b) given the time t to their common ancestor for a reversible mutation rate matrix (such as $\mathbf{Q}$) is

$$P(a,b \mid t) = \delta_{ab}\pi_a p_{ab}^{(2t)}, \tag{7}$$

where $\pi_0 = \pi_1 = 1/2$ are the equilibrium frequencies of states 0 and 1, and

$$\delta_{ab} = \begin{cases} 1 & \text{for } a = b \\ 2 & \text{for } a \neq b \end{cases}.$$

So

$$P(a,b \mid t) = \begin{cases} \frac{1}{4} + \frac{1}{4}\exp\{-\theta t\} & \text{for } a = b \\ \frac{1}{2} - \frac{1}{2}\exp\{-\theta t\} & \text{for } a \neq b \end{cases}.$$

To obtain the probability of observing a pair of states unconditional on the time to their common ancestor involves the integration

$$P(a,b)=\int_0^\infty P(a,b\mid t)P(t)\,\mathrm{d}t\,, \tag{8}$$

where $P(t)=k\exp\{-kt\}$ from before. Therefore the emission probability is defined by

$$P(a,b)=\begin{cases}\dfrac{2k+\theta}{4(k+\theta)} & \text{for } a=b\\ \dfrac{\theta}{2(k+\theta)} & \text{for } a\neq b\end{cases},$$

which is normalised because $P(0,0)+P(0,1)+P(1,1)=1$. Li and Stephens (2003) denote the emission probability

$$\gamma_i(x)=P(H_{k+1,i},H_{x,i}). \tag{9}$$

4.2.2.3 Recombination model

The transmission probability models recombination, because it specifies the probability of a switch from copying one haplotype to copying another between adjacent sites *i* and (*i* + 1). Li and Stephens (2003) model the length of sequence before a switch as exponentially distributed with rate ρ/k. This is based on the informal idea that $E(t)=1/k$, so the average rate of recombination between a pair of sequences is roughly $(\rho/2)\times(2/k)$. Under this crude approximation, the transmission probability is defined by

$$P(X_{i+1}=x'\mid X_i=x)=\begin{cases}\exp\{-\rho_i d_i/k\}+(1-\exp\{-\rho_i d_i/k\})/k & \text{if } x'=x\\ (1-\exp\{-\rho_i d_i/k\})/k & \text{otherwise}\end{cases} \tag{10}$$

where X_i is the copied haplotype at site i, X_{i+1} is the copied haplotype at site $(i + 1)$, and d_i is the distance (in bp) between sites i and $(i + 1)$. In this model there can be a different recombination rate ρ_i between every pair of adjacent sites.

4.2.2.4 Computing the likelihood

To calculate the approximate conditional likelihood requires a summation over all possible combinations of the latent variable at every site; that is to say, all possible mosaics that might constitute the $(k+1)$th haplotype. The advantage of the HMM is that this computation is fast using the forward algorithm (e.g. Rabiner 1989). Suppose that $\alpha_i(x)$ is the joint likelihood of the first i sites and $X_i = x$. Then the approximate conditional likelihood is

$$\hat{\pi}(H_{k+1} \mid H_1, H_2, \ldots, H_k, \Theta) = \sum_{x=1}^{k} \alpha_L(x),$$

when there are L sites. From the forward algorithm,

$$\begin{aligned} \alpha_{i+1}(x) &= \gamma_{i+1}(x) \sum_{x'=1}^{k} \alpha_i(x') P(X_{i+1} = x \mid X_i = x') \\ &= \gamma_{i+1}(x) \left(p_i \alpha_i(x) + (1 - p_i) \frac{1}{k} \sum_{x'=1}^{k} \alpha_i(x') \right), \end{aligned} \tag{11}$$

where $p_i = \exp\{-\rho_i d_i / k\}$. Because the second term in Equation 11 does not depend on x, it only needs to be computed once for each site. As a result, the computational complexity of the approximate conditional likelihood $\hat{\pi}$ is linear in L and linear in the total sample size n. The complexity of the full PAC likelihood is, therefore, linear in L and quadratic in n (Li and Stephens 2003).

4.2.3 NY98 in the coalescent approximation

Incorporating the NY98 mutation model in to the coalescent approximation of Li and Stephens (2003) is straightforward. The instantaneous mutation rate matrix **Q** in Equation 5 is replaced by that defined by Equation 1. However, the exponentiation of the NY98 rate matrix in Equation 6 cannot be solved analytically. Instead, a numerical technique known as diagonalisation is used. Equation 6 can be re-written using the matrix factorisation

$$\mathbf{P}^{(t)} = \mathbf{V}e^{t\mathbf{D}}\mathbf{V}^{-1} \tag{12}$$

(Grimmett and Stirzaker 2001) where **V** is a matrix whose columns are the right eigenvectors of **Q**, $\mathbf{V}^{-1}$ is its inverse and **D** is a diagonal matrix whose diagonal elements are the eigenvalues of **Q**. Exponentiation of a diagonal matrix is trivial, because

$$\exp\{t\mathbf{D}\}_{ij} = \begin{cases} \exp\{d_{ij}t\} & \text{for } i = j \\ 0 & \text{for } i \neq j \end{cases}. \tag{13}$$

So, breaking down Equation 12 into parts for simplification,

$$\mathbf{P}^{(t)} = \mathbf{M}\mathbf{V}^{-1}$$

where

$$\mathbf{M} = \mathbf{V}e^{t\mathbf{D}}.$$

Now, using Equation 13 and the laws of matrix multiplication,

$$m_{ij} = v_{ij}\exp\{d_{jj}t\}$$

so

$$p_{ij}^{(t)} = \sum_{c \in C} v_{ic}\exp\{d_{cc}t\}v_{cj}^{(-1)}, \tag{14}$$

where C is the state space of **Q**, which consists of the 61 non-stop codons for NY98. Using Equation 7, the probability of observing a pair of states $a = H_{k+1,i}$ and $b = H_{x,i}$ when the $(k+1)$th haplotype is copying from the xth haplotype is,

$$P(a,b \mid t) = \delta_{ab}\pi_a \sum_{c \in C} v_{ac} v_{cb}^{(-1)} \exp\{2d_{cc}t\}.$$

Following Equation 8, one can obtain an expression for the HMM emission probability under any reversible mutation matrix **Q**

$$P(a,b) = \delta_{ab}\pi_a \sum_{c \in C} v_{ac} v_{cb}^{(-1)} \frac{k}{k - 2d_{cc}}. \quad (15)$$

Equation 15 is useful because it means that the PAC likelihood can be adapted to any reversible mutation model, of which NY98 is just an example (e.g. Rodríguez *et al.* 1990; Goldman and Yang 1994; Sainudiin *et al.* 2005). For a particular combination of the mutation rate parameters μ, κ and ω, the rate matrix **Q** must be diagonalised, which is to say its eigenvalues and right eigenvectors must be found (Equation 12). This can be achieved for any general real matrix **Q** using a numerical algorithm, available in libraries such as Numerical Recipes (Press *et al.* 2002), LAPACK (Anderson *et al.* 1999) or NAG. See Wilkinson and Reinsch (1971) for details of the algorithm. One problem with the algorithm for diagonalising a general real matrix is that the eigenvalues and eigenvectors are not guaranteed to be real numbers. In fact the eigenvalues and eigenvectors of a reversible rate matrix are real. I am grateful to Ziheng Yang for showing how further factorisation of Equation 12 leads to diagonalisation of a symmetric real matrix, for which the algorithms are guaranteed to produce real eigenvalues and eigenvectors. The algorithm for diagonalising a symmetric real matrix is also quicker and safer than the algorithm for diagonalising a

general real matrix. The code I used for the implementation of this algorithm was kindly provided by Ziheng Yang.

A reversible, irreducible mutation rate matrix **Q**, which is given by Equation 1 for NY98, can be re-written

$$\mathbf{Q} = \mathbf{S\Pi}$$

where **S** is a symmetric matrix ($s_{ij} = q_{ij} / \pi_j, i \neq j$, cf. Equation 1) and **Π** is a diagonal matrix whose diagonal elements are the stationary frequencies π_j of the rate matrix. The eigenvalues and eigenvectors of **Q** can be obtained by constructing a symmetric matrix

$$\mathbf{A} = \mathbf{\Pi}^{1/2}\mathbf{Q}\mathbf{\Pi}^{-1/2},$$

because the eigenvalues of **A** and **Q** are the same (contained in the diagonal matrix **D**), and the matrix of right eigenvectors **V** for matrix **Q** is related to the matrix of right eigenvectors **R** for matrix **A** by the formulae

$$\mathbf{V} = \mathbf{\Pi}^{-1/2}\mathbf{R},$$
$$\mathbf{V}^{-1} = \mathbf{R}^{-1}\mathbf{\Pi}^{1/2}.$$

Matrices **D** and **R** are obtained by diagonalising **A** using the algorithm for a symmetric real matrix. Because **R** is orthogonal, $\mathbf{R}^{-1} = \mathbf{R}^{T}$, so no matrix inversion is required for obtaining $\mathbf{V}^{-1}$. By matrix multiplication

$$\begin{aligned} v_{ac} &= \pi_a^{-1/2} r_{ac} \\ v_{cb}^{(-1)} &= r_{bc} \pi_b^{1/2} \end{aligned}. \qquad (16)$$

Therefore, Equation 15 can be re-written

$$P(a,b) = \delta_{ab} \pi_a^{1/2} \pi_b^{1/2} \sum_{c \in C} r_{ac} r_{bc} \frac{k}{k - 2d_{cc}}. \qquad (17)$$

This is the actual formula used in the implementation of the model.

4.2.4 An indel model for NY98

Alignments of nucleotide sequences from antigen loci are punctuated by gaps in the alignment caused by insertion or deletion mutations (indels). A sequence alignment is a statement of the homology of particular nucleotides in one sequence to those in the other sequences. Indels cause gaps in the nucleotide sequence alignment in multiples of three when the gene is functional, because otherwise a frameshift will ensue, and the remaining sequence will be nonsense. Indels are an important feature of the evolution of antigen loci, but even simple treatments of indels result in complex models that do not share the nice properties of the reversible nucleotide and codon models in common usage (e.g. Thorne *et al.* 1991, 1992). Here I make a very simple extension of NY98 in order to incorporate an extra indel state. The motivation for using this model is not to provide a realistic model of insertion/deletion, but to capture the information regarding the underlying tree structure and mode of selection at sites segregating for indels in the simplest possible way. The model is only applied to columns in the alignment that are segregating for an indel.

For columns segregating for an indel, codons are assumed to mutate to the indel state at rate $\pi_{\text{indel}}\varphi\omega$ and back at rate $(1-\pi_{\text{indel}})\varphi\omega$. Here π_{indel} is the equilibrium frequency of indels (in sites segregating for indels), φ is the rate of insertion/deletion, and ω is the selection parameter for that site. The model can be thought of in two parts: the NY98 model is nested within a two state codon vs. indel model (0 = codon, 1 = indel) specified by

$$\mathbf{Q}^{*} = \begin{vmatrix} -\pi_{\text{indel}}\varphi\omega & \pi_{\text{indel}}\varphi\omega \\ (1-\pi_{\text{indel}})\varphi\omega & -(1-\pi_{\text{indel}})\varphi\omega \end{vmatrix}. \quad (18)$$

Exponentiating Equation 18 gives the transition probability matrix between codon and indel states. So

$$p_{ij}^{*(t)} = \begin{cases} 1-\pi_{indel}(1-\exp\{-\omega\varphi t\}) & \text{for two (unspecified) codons} \\ \pi_{indel}+(1-\pi_{indel})\exp\{-\omega\varphi t\} & \text{for two indels} \\ \pi_{indel}(1-\exp\{-\omega\varphi t\}) & \text{if i is a codon and j an indel} \\ (1-\pi_{indel})(1-\exp\{-\omega\varphi t\}) & \text{if i is an indel and j a codon} \end{cases}. \quad (19)$$

Denote the full transition probability matrix for the NY98 model with indels $\mathbf{P}^{(t)}$. From Equation 19, part of this matrix is apparent

$$p_{ij}^{(t)} = \begin{cases} \pi_{indel}+(1-\pi_{indel})\exp\{-\omega\varphi t\} & \text{for two indels} \\ \pi_{indel}(1-\exp\{-\omega\varphi t\}) & \text{if i is a codon and j an indel} \\ \pi_{j}(1-\pi_{indel})(1-\exp\{-\omega\varphi t\}) & \text{if i is an indel and j a codon} \end{cases}.$$

When *i* and *j* are both codons, $p_{ij}^{(t)}$ can be found by conditioning on whether there are intermediate indels. Denote $\mathbf{N}^{(t)} = \{\nu_{ij}^{(t)}\}$ for the transition probability matrix of the NY98 model without indels. Conditional on intermediate indels, the transition probability from codon *i* to *j* in time *t* is simply π_j. Conditional on no intermediate indels, the transition probability from codon *i* to *j* in time *t* is $\nu_{ij}^{(t)}$. Since the probability of no intermediate indels is $\exp\{-\pi_{\text{indel}}\varphi\omega\}$, for a pair of codons

$$p_{ij}^{(t)} = \nu_{ij}^{(t)}\exp\{-\pi_{\text{indel}}\varphi\omega\} + \pi_j\left[1-\pi_{indel}(1-\exp\{-\omega\varphi t\}) - \exp\{-\pi_{\text{indel}}\varphi\omega\}\right].$$

Using Equations 8, 14 and 16 the emission probabilities for the PAC likelihood are obtained. For two identical codons

$$P(a,a) = \pi_a^2(1-\pi_{indel})\left[(1-\pi_{indel}) + \frac{k\pi_{indel}}{k+2\omega\varphi} - \frac{k}{k+2\pi_{indel}\omega\varphi} + \frac{1}{\pi_a}\sum_{c\in C}\frac{kr_{ac}r_{bc}}{k+2\pi_{indel}\omega\varphi-2d_{cc}}\right] \quad (20a)$$

where *C* is the state space of the NY98 model. For two non-identical codons

$$P(a,b) = 2\pi_a\pi_b(1-\pi_{indel})\left[(1-\pi_{indel}) + \frac{k\pi_{indel}}{k+2\omega\varphi} - \frac{k}{k+2\pi_{indel}\omega\varphi} + \frac{1}{\sqrt{\pi_a\pi_b}}\sum_{c\in C}\frac{kr_{ac}r_{bc}}{k+2\pi_{indel}\omega\varphi-2d_{cc}}\right] \quad (20b)$$

For two indels

$$P(a,a) = \pi_{indel}^2 + \frac{\pi_{indel}k(1-\pi_{indel})}{k+2\omega\varphi}. \quad (20c)$$

For a codon *a* and an indel *b*

$$P(a,b) = 2\pi_a\pi_b(1-\pi_{indel})\left(1-\frac{k}{k+2\omega\lambda}\right). \quad (20d)$$

4.2.5 Variation in ω and ρ along a gene

The primary aim of the new method is to obtain posterior distributions for ω and ρ, allowing both to vary along the length of the sequence. The information regarding either ω or ρ at a given position along the sequence is limited by the number of mutations in the underlying evolutionary history. This is a potentially serious limitation, particularly for sequences with low diversity. In an attempt to exploit to the full the available information, I use a independent prior distributions on ω and ρ in which adjacent sites may share either parameter in common. I will describe the model of variation in ω for the purposes of information. The model of variation for ρ is of the same form.

For a sequence of length L codons, the prior distribution imposes a 'block-like' structure on the variation in ω with two fixed and B_ω $(0 \le B_\omega \le L-1)$ variable transition points,

$$\mathbf{s}^{(B_\omega)} = (s_0, s_1, \ldots, s_{B_\omega+1}),$$

where $(s_0 = 0) < s_1 < s_2 < \cdots < s_{B\omega} < (s_{B_\omega+1} = L).$

Block j is delimited by transition points (s_j, s_{j+1}) and has a common selection parameter ω_j. I model the number of variable transition points in the region as a binomial distribution with parameters $(L-1, p_\omega)$. Given the number of transition points, the selection parameter for each block is independently and identically distributed. For an exponential prior on ω_j with rate parameter λ, the prior distribution on the transition points and selection parameters can be written

$$P\left(B, \mathbf{s}^{(B_\omega)}, \boldsymbol{\omega}^{(B_\omega)}\right) = p_\omega^{B_\omega} (1-p_\omega)^{L-B_\omega-1} \lambda^{B_\omega+1} \exp\left\{-\lambda(\omega_0 + \omega_1 + \cdots + \omega_{B_\omega})\right\} \quad (21)$$

In this model, the expected length of a block is $L/([L-1]p_\omega + 1) \approx 1/p_\omega$. For $p_\omega = 0$ there is a single block, producing a constant model for ω along the sequence, and for $p_\omega = 1$ every site has its own independent ω.

This prior structure is based on the multiple change-point model of Green (1995) which was adopted by McVean *et al*. (2004) to estimate variable recombination rates along a gene sequence, although the binomial model that I have used here is designed specifically so that transition points must fall between codons at a finite $(L-1)$ number of positions. I implement a block-like prior on ρ of the same form as for ω, but the block structure for ρ is independent of the block structure for ω, and the number of variable transition points is binomially distributed with parameters $(L-2, p_\rho)$. It is assumed that recombination only occurs between codons and not within. To perform inference jointly on variation in ω and ρ along the sequence I will use reversible-jump MCMC.

Table 1 Notation used for Constants

n	Sample size
L	Number of codons
P	Ploidy
N_e	Effective population size

Table 2 Parameters of the Model

μ	Rate of synonymous transversion per PN_e generations
κ	Transition:transversion ratio
B_ω	Number of changes in the dN/dS ratio along the sequence
$s_j^{(B_\omega)}, j = 0 \ldots B_\omega + 1$	Positions at which the dN/dS ratio changes along the sequence
$\omega_j, j = 0 \ldots B_\omega$	dN/dS ratio between sites $s_j^{(\omega)}$ and $s_{j+1}^{(\omega)}$
B_ρ	Number of changes in the recombination rate along the sequence
$s_j^{(B_\rho)}, j = 0 \ldots B_\rho + 1$	Positions at which the recombination rate changes along the sequence
$\rho_j, j = 0 \ldots B_\rho$	Recombination rate between sites $s_j^{(\rho)}$ and $s_{j+1}^{(\rho)}$
φ	Rate of insertion/deletion per PN_e generations

4.3 Bayesian inference

To summarise, Tables 1 and 2 list the constants and parameters of the model. The parameters together in Table 2 are denoted Θ, and the aim of Bayesian inference is to obtain a posterior distribution of Θ given the data **H**. To do so I will use Markov Chain Monte Carlo (MCMC; see for example O'Hagan and Forster [2004] for

Table 3 MCMC Moves

Type	Move	Relative proposal probability
A	Change μ	●●
A	Change κ	●●
A	Change φ	●●
A	Change ω within a block	●●●
A	Change ρ within a block	●●●
B	Extend an ω block 5′ or 3′	●●●
B	Extend an ρ block 5′ or 3′	●●●
C/D	Split or merge ω blocks	●●●●●●
C/D	Split or merge ρ blocks	●●●●●●

details). In brief, the Markov chain is initiated using values taken at random from the priors. Each iteration of the chain one or more parameters are updated according to a proposal distribution, and the proposal is accepted with the acceptance probabilities specified in the next section. There are nine moves that can be proposed, each of which is visited with the relative probability specified in Table 3. This is known as a random sweep. Moves of type A and B (Table 3) are Metropolis-Hastings (Metropolis *et al.* 1953; Hastings 1970) moves that change a single parameter at a time. Moves of type C and D are complementary reversible-jump moves (Green 1995). For the purpose of illustration, I will describe one each of move types A-D, and assume that the prior on the ω_j's specifies i.i.d. exponential distributions with rate λ. The moves below describe in full how variation in ω along the sequence is explored by MCMC.

4.3.1 Type A. Change ω within a block

Metropolis-Hastings move

A block is chosen uniformly at random. A new value ω' is proposed so that $\omega' = \omega \exp(U)$ where U ~ Uniform(-1,1). Thus $\omega e^{-1} < \omega' < \omega e$. The acceptance probability is given by the Metropolis-Hastings ratio

$$\alpha(\Theta \to \Theta') = \min\left\{1, \frac{P(\mathbf{H} \mid \Theta')}{P(\mathbf{H} \mid \Theta)} \frac{P(\Theta')}{P(\Theta)} \frac{K(\Theta' \to \Theta)}{K(\Theta \to \Theta')}\right\},$$

where $K(\Theta \to \Theta')$ is the proposal kernel density. To find *K*, note that

$$\Pr(U < u) = \frac{1}{2}(1+u), \ -1 < u < 1.$$

So

$$\begin{aligned} \Pr(\omega' < x) &= \Pr(\omega e^{U} < x) \\ &= \Pr\left(U < \ln\frac{x}{\omega}\right) \\ &= \frac{1}{2}\left(1 + \ln\frac{x}{\omega}\right). \end{aligned}$$

Therefore

$$\begin{aligned} P(\omega' = x) &= \frac{\partial}{\partial x} \Pr(\omega' < x) \\ &= \frac{1}{2x}. \end{aligned}$$

This gives an acceptance probability of

$$\alpha_A(\Theta \to \Theta') = \min\left\{1, \frac{P(\mathbf{H} \mid \Theta')}{P(\mathbf{H} \mid \Theta)} \exp\{-\lambda(\omega' - \omega)\} \frac{\omega'}{\omega}\right\}. \qquad (22)$$

4.3.2 Type B. Extend an ω block 5′ or 3′

Metropolis-Hastings move

The block to extend is chosen uniformly at random, and for each block the direction is chosen with equal probability. If the 5′-most or 3′-most block is chosen to be extended 5′ or 3′ respectively, the move is rejected. The number of sites to extend the block, $g \in [1, \infty)$ is chosen from a geometric distribution with some parameter. If extending the block *g* sites in the chosen direction would cause it to merge with the adjacent block, the move is rejected.

The proposal distribution is symmetric, so the Hastings ratio is one. The ratio of priors is also one because the prior on the positions of the transition points is uniform. Therefore

$$\alpha_B(\Theta \to \Theta') = \min\left\{1, \frac{P(\mathbf{H} \mid \Theta')}{P(\mathbf{H} \mid \Theta)}\right\}. \tag{23}$$

4.3.3 Types C and D. Split and Merge an ω block

Reversible Jump moves

The acceptance probability for a reversible jump move (Green 1995) is

$$\alpha_m(\Theta \to \Theta') = \min\left\{1, \frac{P(\mathbf{H} \mid \Theta')}{P(\mathbf{H} \mid \Theta)} \frac{P(\Theta')}{P(\Theta)} \frac{j_m(\Theta')}{j_m(\Theta)} \frac{g'_m(\mathbf{U}')}{g_m(\mathbf{U})} \left|\frac{\partial(\Theta', \mathbf{U}')}{\partial(\Theta, \mathbf{U})}\right|\right\}.$$

Here $j_m(\Theta)$ is the probability of proposing move *m* when at state Θ, and $g_m(\mathbf{U})$ is the joint probability density of the random vector **U** which is generated to facilitate

the transformation from $(\Theta, \mathbf{U})$ to $(\Theta', \mathbf{U}')$. The last term in the acceptance probability is the determinant of the Jacobian of the diffeomorphism (the transformation which must be differentiable in both directions).

4.3.3.1 Ratio of priors

In move C a block that currently has length $(s_{j+1} - s_j)$ is split at position *s**, and its current selection parameter ω_j is transformed, with the aid of a random variable *U*, into two new parameters ω'_j and ω'_{j+1}. The ratio of priors is

$$\frac{p_\omega}{(1-p_\omega)} \lambda \exp\{-\lambda(\omega'_j + \omega'_{j+1} - \omega_j)\}.$$

In move D two adjacent blocks that currently have lengths $(s^* - s_j)$ and $(s_{j+1} - s^*)$ are merged, and their selection parameters ω_j and ω_{j+1} are transformed into a single parameter ω'_j. So the ratio of priors is

$$\frac{(1-p_\omega)}{p_\omega \lambda} \exp\{-\lambda(\omega'_j - \omega_j - \omega_{j+1})\}.$$

4.3.3.2 Ratio of proposal probabilities

Move C splits an existing block. When there are $(B_\omega + 1)$ blocks there are $(L - B_\omega - 1)$ possible positions at which a block could be broken. The position of the split, *s**, is chosen uniformly at random from these. Move type C_i splits the block that spans position *i*; only $(L - B_\omega - 1)$ out of the total possible $L - 1$ type C moves are

available at any one time. So $j_{C_i}(\Theta) = c_B/(L - B_\omega - 1)$, where c_B is the total rate at which type C moves are proposed when there are $(B_\omega + 1)$ blocks.

Move D merges two adjacent blocks. Assuming that the block merges with its 3′ neighbour, there are B_ω possible mergers. The merger is chosen uniformly at random from these B_ω possibilities. So $j_{D_i}(\Theta) = d_B/B_\omega$, where d_B is the total rate at which type D moves are proposed when there are $(B_\omega + 1)$ blocks.

Following Green (1995), when there are B_ω transition points, moves C and D are proposed with relative probabilities c_B and d_B, where

$$\frac{c_B}{d_B} = \frac{\min\{1, P(B_\omega + 1)/P(B_\omega)\}}{\min\{1, P(B_\omega - 1)/P(B_\omega)\}}.$$

Under the prior, the number of transition points B_ω is distributed binomially. This yields

$$\frac{\Pr(B_\omega + 1)}{\Pr(B_\omega)} = \frac{(L - B_\omega - 1)}{(B_\omega + 1)} \frac{p_\omega}{(1 - p_\omega)} \text{ and } \frac{\Pr(B_\omega)}{\Pr(B_\omega - 1)} = \frac{B_\omega}{(L - B_\omega)} \frac{(1 - p_\omega)}{p_\omega}.$$

4.3.3.3 Ratio of density functions

In transforming ω_j to ω'_j and ω'_{j+1}, it is necessary to introduce a random deviate U to match the dimensionality on both sides. So the transformation $(\omega_j, U) \to (\omega'_j, \omega'_{j+1})$ involves the generation of a random deviate U in move C, but not in the inverse move D. This simplifies $g_D(\mathbf{U}')/g_C(\mathbf{U})$ to $1/g_C(U)$. Since U is chosen uniformly on $(0,1)$, this ratio equals one.

4.3.3.4 Jacobian

In Move C the values of the selection parameters for the two resulting blocks, ω'_j and ω'_{j+1} are chosen from the current value of ω_j so that the weighted geometric mean is preserved. The weighting takes into account the relative sizes of the two resulting blocks, which are $(s*-s_j)$ and $(s_{j+1}-s*)$ respectively. Thus

$$\omega_j'^{(s*-s_j)}\omega_{j+1}'^{(s_{j+1}-s*)}=\omega_j^{(s_{j+1}-s_j)}.$$

To introduce a random element,

$$\frac{\omega'_{j+1}}{\omega'_j}=\frac{1-U}{U},$$

where U ~ Uniform(0,1). The determinant of the Jacobian is,

$$J=\begin{vmatrix}\dfrac{\partial\omega'_j}{\partial\omega_j} & \dfrac{\partial\omega'_{j+1}}{\partial\omega_j}\\ \dfrac{\partial\omega'_j}{\partial U} & \dfrac{\partial\omega'_{j+1}}{\partial U}\end{vmatrix},$$

To obtain J, it is necessary to express ω'_j and ω'_{j+1} in terms of ω_j and U, giving

$$\omega'_j=\omega_j\left(\frac{U}{1-U}\right)^{1-a}$$

and

$$\omega'_{j+1}=\omega_j\left(\frac{1-U}{U}\right)^{a},$$

where $a=(s*-s_j)/(s_{j+1}-s_j)$. The determinant of the Jacobian (which is defined to be always positive) comes out as

$$J=\frac{(\omega'_j+\omega'_{j+1})^2}{\omega_j}.$$

4.3.3.5 Acceptance probabilities

For move C,

$$\alpha_C(\Theta \to \Theta') = \min\left\{1, \frac{P(\mathbf{H} \mid \Theta')}{P(\mathbf{H} \mid \Theta)} \frac{p_\omega \lambda \mathrm{e}^{-\lambda(\omega'_j + \omega'_{j+1})}}{(1-p_\omega)\mathrm{e}^{-\lambda(\omega_j)}} \frac{d_{B_\omega+1}(L - B_\omega - 1)}{c_{B_\omega}(B_\omega + 1)} \frac{(\omega'_j + \omega'_{j+1})^2}{\omega_j}\right\}. \quad (24)$$

For move D,

$$\alpha_D(\Theta \to \Theta') = \min\left\{1, \frac{P(\mathbf{H} \mid \Theta')}{P(\mathbf{H} \mid \Theta)} \frac{(1-p_\omega)\mathrm{e}^{-\lambda(\omega'_j)}}{p_\omega \lambda \mathrm{e}^{-\lambda(\omega_j + \omega_{j+1})}} \frac{c_{B_\omega - 1} B_\omega}{d_{B_\omega}(L - B_\omega)} \frac{\omega'_j}{(\omega_j + \omega_{j+1})^2}\right\}. \quad (25)$$

Table 4 Structure of the *omegaMap* program

File	Function	# Lines
main.h	Header file for main.cpp	6
omegaMap.h	Header file for omegaMap.cpp	361
main.cpp	Program control	30
omegaMap.cpp	Read in command line and configuration file options. Allocate memory. Initialize the MCMC chain.	1164
likelihood.cpp	Calculate the likelihood. Forward and backward algorithm. Build the mutation rate matrix.	726
mcmc.cpp	Controls the MCMC scheme. Proposes moves. Calculates acceptance probabilities.	1514
io.cpp	Outputs MCMC chain in text format and encoded format. Functions for reading in MCMC chain from encoded format.	504

Table 5 Utilities used by *omegaMap*

File	Function	# Lines
argumentwizard.h	Utility for reading in command line options.	215
controlwizard.h	Utility for reading in configuration files.	659
dna.h	Functions for reading in FASTA files and storing DNA sequences.	486
lotri_matrix.h	Lower triangular matrix class.	144
matrix.h	Matrix class.	226
myerror.h	Error and warning functions.	33
myutils.h	Links these various utility files.	35
random.h	Random number generation.	520
utils.h	Various utilities.	29
vector.h	Vector class.	133

Table 6 *PAML* package, linked to by *omegaMap*

File	Function	# Lines
paml.h	Header file for tools.c	335
tools.c	PAML functions	4369

PAML was written by Ziheng Yang and is available from

http://abacus.gene.ucl.ac.uk/software/paml.html

4.3.4 Implementation

I implemented the likelihood calculation and inference scheme in C++. The program, called *omegaMap*, was built up progressively, from testing the likelihood function on simple examples that could be verified using a calculator, to a Metropolis-Hastings MCMC scheme without variation in ω and ρ, to the full reversible-jump MCMC scheme. The code was developed using *Microsoft Visual C++* and then switched to Linux *gcc* for testing on datasets of realistic size. The MCMC scheme was debugged principally by using a flat likelihood, in which case one expects to recover the prior from the posterior. This proved important when, having moved from a dual-node 64-bit AMD machine (mcv1@stats.ox.ac.uk) I recompiled the program on a multi-node 64-bit AMD machine (genecluster@stats.ox.ac.uk), the posterior began to produce a systematic bias in the recombination rate estimates, so that rates declined 5′-3′, even when the same sequence was reversed. Using a flat likelihood revealed that there was a numerical inconsistency, probably caused by a difference in compilers on the two machines. The problem was solved in a makeshift fashion by running the executable compiled on mcv1 on genecluster. This was a compromise because the executable compiled on mcv1 ran somewhat slower on genecluster than the executable compiled on genecluster. This is a cause for concern because the expectation is that C++ code is portable between machines and compilers. As a result when the code is distributed I will stress the need to test the program by compiling it first with flat likelihoods (which can be done using the flag `-D _TESTPRIOR`) and ensuring the prior is recovered from the posterior.

Table 7 Structure of the *analyse* program

File	Function	# Lines
analyse.h	Header file for analyse.cpp	45
main.cpp	Program control	73
analyse.cpp	Functions for reconstructing the MCMC chain based on an encoded file.	411

Tables 4-6 show the structure of the *omegaMap* program. In total there are 6,785 lines of novel code (Tables 4 and 5). *omegaMap* uses some functions in the *PAML* package (Table 6), written by Ziheng Yang. *PAML* (Phylogenetic Analysis by Maximum Likelihood) is freely available from http://abacus.gene.ucl.ac.uk/software/paml.html. In addition, many functions in the C++ standard template library are used, so the total size of the code is unknown. *omegaMap* can output the results in two formats. The first is a tab-delimited text file with a column for each parameter in the model and a number of other diagnostics such as the acceptance probability and computational time. The thinning interval dictates the number of iterations before the parameter state is output. This text file can be read by software such as *R* or *Excel*. However, outputting the entire MCMC chain using a thinning interval of one creates an enormous text file with a great deal of redundancy because only a subset of the parameters are changed in any iteration. Therefore *omegaMap* can output in a second format, an encoded version of the MCMC chain. The program *analyse* (Table 7) can read this file, reconstruct the MCMC chain internally (orders of magnitude faster than the original MCMC chain was generated) and output a text file for use with *R* or *Excel*.

4.4 Simulation study

To investigate the performance of the method, I undertook two simulation studies. In one data was simulated with variation in the selection parameter along the sequence, and a constant recombination rate. In the other, data was simulated with variation in the recombination rate along the sequence, and a constant selection parameter. Each study consisted of simulating 100 datasets of $n = 20$ sequences each of length $L = 200$ codons using the coalescent with recombination (Hudson 1983, Griffiths and Marjoram 1997) and the NY98 mutation model. Every simulated dataset was analysed twice, using 250,000 iterations of the MCMC and a burn-in of 20,000 iterations. Initial values were chosen randomly from the priors independently for the two runs. The runs were compared for convergence and merged to obtain the posterior distributions.

4.4.1 Permutation test for recombination

Before the datasets were analysed, each was subjected to a permutation test for recombination (McVean *et al.* 2001; Meunier and Eyre-Walker 2001). Phylogenetic analysis is inappropriate for gene sequences taken from populations that are demonstrably recombinogenic. The aim of the permutation tests was to demonstrate the recombinogenic nature of the data.

The permutation test is a goodness-of-fit test for the model of no recombination. When there is no recombination, there ought to be no correlation between physical distance and LD, so sites are exchangeable. It should be noted that sites are also exchangeable in the case of complete linkage equilibrium. If LD tails off with

physical distance then recombination must have occurred in the ancestral history of the sequences. The test proceeds as follows

1. The observed correlation between a measure of LD and physical distance is recorded as c_{obs}.
2. The nucleotide positions are reordered at random and the correlation between LD and physical distance is calculated.
3. Step 2 is repeated 999 times.

Three measures of LD can be used: r^2 (Hill and Robertson 1968), D' (Lewontin 1964) and the four-gamete test (*G*4; Hudson and Kaplan 1985). In section 2.3.2 cor(r^2,*d*), where *d* is physical distance, was used for testing the goodness-of-fit of the standard neutral coalescent. If c_{obs} lies in the tail of the reference distribution then the model of exchangeability of sites is not a good fit to the data, and we can conclude that there is good evidence for recombination in the data. The probability of obtaining a result as extreme as observed under the model can be expressed as a *p* value, where *p* is estimated to be

$$p = \frac{n+1}{N+1}$$

(Sokal and Rohlf 1995). Here *n* is the number times a value more extreme than c_{obs} was observed out of a total of *N* simulations.

Using *p* values to reject a "null" model might seem to be a particularly frequentist thing to do. In fact a frequentist *p* value and a Bayesian posterior predictive *p* value (Rubin 1984) are equivalent in the model of exchangeability described here, because the model has no parameters. I will discuss the use of posterior predictive *p* values for goodness-of-fit testing more in chapter 5.

4.4.2 Simulation study A

This study was designed to simulate data with variation in ω but not in ρ. I varied ω between 0.1 and 10, as shown by the red line in Figure 5a. I created more fine detail in variation in ω for $\omega > 1$ because, biologically, a scenario in which there is an excess of non-synonymous relative to synonymous polymorphism is of greater interest. For the same reason ω is plotted on a natural, rather than a logarithmic scale. The mutation parameters were set at $\mu = 0.7$ and $\kappa = 3.0$, which gives $\theta_S = 0.1$. The recombination rate was set constant at $\rho = 0.1$, giving a total recombination distance for the region of $R = \sum \rho = 19.9$. The mutation and recombination parameters were chosen to mimic those estimated for the housekeeping genes of *Neisseria meningitidis* (see Chapter 1). Exponential distributions were used for the priors on μ, κ, ω and ρ, with means 0.7, 3.0, 1.0 and 0.1.

Permutation tests showed that phylogenetic analysis of these datasets was inappropriate because of the presence of recombination. The number of datasets for which the *p*-value was less than 0.05 was 99, 93 and 93 for the three measures of LD (r^2, D' and *G*4) respectively.

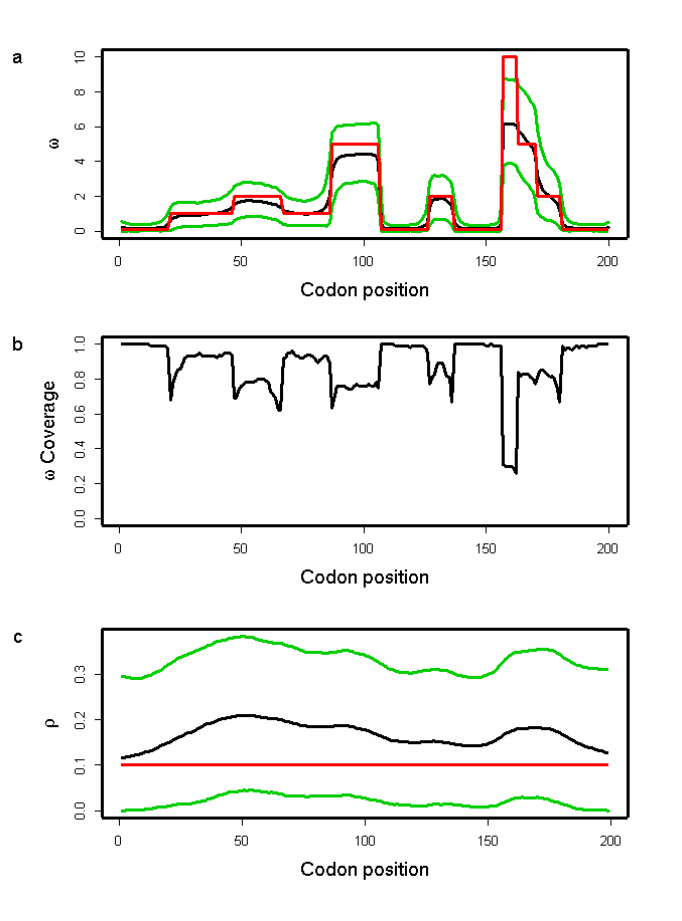

Figure 5 Results of simulation study A. (a) Average posterior of ω, (b) coverage of ω and (c) average posterior of ρ. In (a) and (c) the red line indicates the truth, the black line indicates the average mean of the posterior and the green lines indicate the average 95% HPD interval of the posterior. The averages are taken over 100 simulated datasets. In (b) coverage is defined as the proportion of the 100 datasets for which the 95% HPD interval encloses the truth.

Figure 5a shows the average over the 100 simulated datasets of the mean and 95% highest posterior density (HPD) interval for the posterior distribution of ω at each site. The average mean posterior density follows the truth closely. Likewise the average 95% HPD interval generally encloses the true value of ω. As expected, the effect of fitting a prior with mean 1 was to cause the posterior to underestimate ω when $\omega > 1$

Table 8 Summary of posteriors for simulation study A

Parameter	Truth	Prior	Average posterior			Coverage
		Mean	Lower 95% HPD	Mean	Upper 95% HPD	
μ	0.7	0.7	0.7	0.9	1.1	0.63
κ	3.0	3.0	2.3	3.1	3.9	0.91
R	19.9	19.9	22.4	33.3	44.7	0.43

and overestimate ω when $\omega < 1$. The effect is not great except for the most extreme values where $\omega = 10$.

However, even where the average 95% HPD interval encloses the truth, that does not mean the 95% HPD interval encloses the truth for all simulated datasets. Figure 5b shows the relevant quantity, the coverage of ω, for each site. Coverage is defined here as the proportion of datasets for which the 95% HPD interval encloses the truth. Half of sites have coverage better than 93%, and 95% of sites have coverage better than 66%. If a false positive is defined as the lower bound of the 95% HPD interval exceeding 1 when in truth $\omega \leq 1$, then the false positive rate was 0.5%. The estimate of the synonymous transversion rate μ exhibits upward bias (average 0.90), with 63% coverage (Table 8), and the transition-transversion ratio κ is estimated to be 3.1 on average, with 91% coverage.

Consistent with the findings of Li and Stephens (2003), I observed that the recombination rate estimator has a small upward bias (Figure 5c). The average mean posterior is almost flat, and the average 95% confidence intervals enclose the truth completely, suggesting that the estimator is good notwithstanding its bias. The

Table 9 MCMC Moves Acceptance Probabilities

Type	Move	Mean acceptance probability α
A	Change μ	0.139
A	Change κ	0.157
A	Change ω within a block	0.573
A	Change ρ within a block	0.727
B	Extend an ω block 5′ or 3′	0.403
B	Extend an ρ block 5′ or 3′	0.825
C	Split an ω block	0.381
D	Merge ω blocks	0.242
C	Split a ρ block	0.635
D	Merge ρ blocks	0.660

coverage is almost constant across sites at 95%. Table 8 shows that the estimate of the total recombination distance, *R*, is also upwardly biased. Coverage of *R*, however, was only 43%, suggesting that the good coverage for ρ at individual sites may be in part because of poor information. Importantly, Figures 5a and 5b show that the effect of the selection parameter on the estimate of ρ is negligible, indicating that inference on ρ is not confounded by ω.

4.4.3 Mixing properties of reversible jump moves

Achieving satisfactory acceptance probabilities can be an issue in reversible-jump MCMC (Green 1995). This was not found to be a problem in the MCMC scheme

presented here. For illustrative purposes, Table 9 shows the acceptance probabilities for the MCMC moves, averaged over a pair of independent analyses of the same dataset from simulation study A. The reversible-jump moves (those of type C or D) had high acceptance probabilities (for example, $\alpha = 0.381$ when splitting an ω block and $\alpha = 0.242$ when merging ω blocks). Of the other moves, acceptance probabilities ranged from 0.139 to 0.825. The lowest acceptance probabilities were for moves changing μ and κ ($\alpha = 0.139$ and 0.157 respectively), perhaps because these changes affect all sites in the sequence unlike any other move. Changes to moves involving ρ had high acceptance probabilities ($\alpha = 0.635$ to 0.825), which may be indicative of the low information regarding variation in recombination rate within the region.

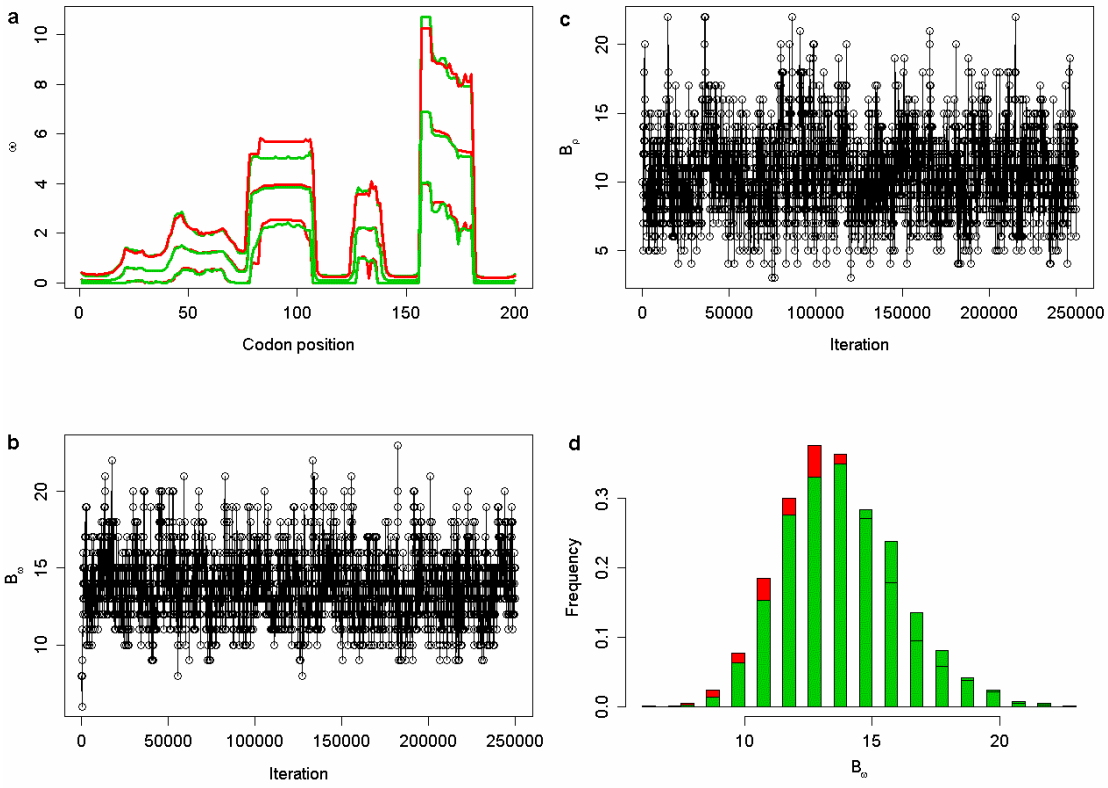

Figure 6 a Convergence of the mean and upper and lower 95% HPD bounds of the posterior on ω for two analyses (red and green lines) of the same dataset from simulation study A. **b** Trace of B_ω for one of the two analyses. **c** Trace of B_ρ for one of the analyses. **d** Convergence of the posterior distribution of B_ω for the two analyses (red and green histograms).

In Figure 6 the mixing properties of the two chains for the same dataset are shown. Figure 6a shows the convergence of the two chains for the posterior distribution on ω across sites. The mean and upper and lower 95% HPD bounds are indicated. One chain is plotted in red, the other in green. The agreement is good; more so for the mean than the 95% HPD bounds. One would expect estimates of the latter to have greater variance. Figure 6b is a trace of B_ω through iterations of one of the Markov chains, and 6c is the corresponding trace of B_ρ. B_ω and B_ρ can only be changed by reversible-jump moves. There is no evidence of poor mixing in either of the traces. Figure 6d shows a histogram of the posterior distribution of B_ω for both the chains

(one in red, the other green). The two appear to converge well throughout the distribution. When the chains are merged the variance in the estimate of the posterior will be reduced. However, if this were an analysis of a real dataset of special interest, rather than one of a hundred simulated datasets, then there is some argument for running the two chains longer to further improve convergence.

4.4.4 Simulation study B

This study was designed to simulate data with variation in ρ but not in ω. Along the sequence ρ was allowed to vary at 0.005, 0.1, 0.5 and 1, for which one would expect 0.018, 0.35, 1.8 and 3.5 recombination events respectively per site in the ancestral history under a coalescent model (Griffiths and Marjoram 1997). The total recombination distance was $R = 37.5$. I let $\mu = 3.6$ and $\kappa = 3.0$ giving $\theta_S = 0.5$, and a constant selection parameter of $\omega = 0.2$. Exponential distributions were used for the priors on μ, κ, ω and ρ, with means 3.6, 3.0, 1.0 and 0.2.

Permutation tests showed that these datasets were not amenable to phylogenetic analysis because of the presence of recombination. All 100 datasets yielded *p*-values less than 0.05 for all three measures of LD.

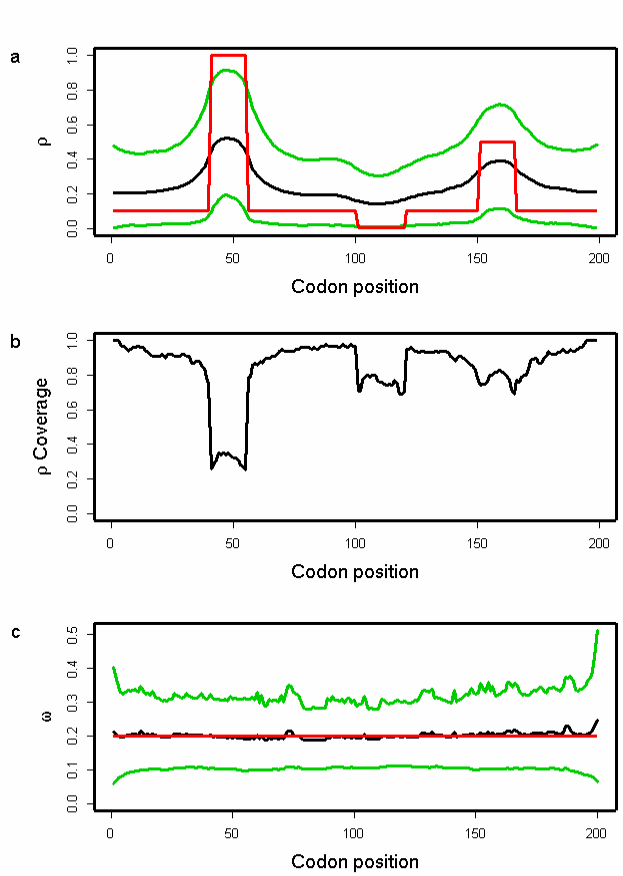

Figure 7 Results of simulation study B. (a) Average posterior of ρ, (b) coverage of ρ and (c) average posterior of ω. In (a) and (c) the red line indicates the truth, the black line indicates the average mean of the posterior and the green lines indicate the average 95% HPD interval of the posterior. The averages are taken over 100 simulated datasets. In (b) coverage is defined as the proportion of the 100 datasets for which the 95% HPD interval encloses the truth.

Variation in the recombination rate was detected by the new method, as seen in Figure 7a. The average over the 100 datasets shows that the mean and 95% HPD interval for the posterior distribution of ρ at each site pick up the rate variation, but not to the full extent. As a result, the coverage shown in Figure 7b is generally good, on average 85%, but performs worst for the most extreme peak in rate between sites 41 and 55, where it consistently underestimates the height. The properties of the estimate of the

Table 10 Summary of posteriors for simulation study B

Parameter	Truth	Prior	Average posterior			Coverage
		Mean	Lower 95% HPD	Mean	Upper 95% HPD	
μ	3.6	3.6	3.4	4.2	5.1	0.53
κ	3.0	3.0	2.5	3.1	3.8	0.95
R	37.5	39.8	37.4	50.9	65.0	0.49

total recombination distance R (Table 10) are similar to those in simulation study A. There is a tendency to overestimate (average 50.9) and as a result coverage is 49%. This bias could be corrected empirically, as in Li and Stephens (2003). Nevertheless, there is power to detect rate variation on such fine scales. The extent to which the posteriors underestimate the deviations from the mean recombination rate reflects the constraining effect of the prior when the signal in the data is weak.

Figure 7c shows that on average the estimates of ω are very close to the truth, with the average 95% HPD intervals completely enclosing the true value. Along the sequence, the estimates are flat, with mean 0.21 and coverage 90%. The false positive rate was zero. Reflecting simulation study A, there was no evidence that variation in the recombination rate confounded inference on the selection parameter. Table 10 shows that there was some upward bias in the mean estimate of $\mu = 4.1$, with 58% coverage, and the transition-transversion ratio was estimated to be 3.2 on average, with 89% coverage. Most importantly, both simulation studies show that when there is variation in ω or ρ it can be detected, when there is no variation none is detected, and there is little or no confounding between ω and ρ.

4.5 Summary

In this chapter I have described a new model for detecting immune selection in nucleotide sequences, based on an approximation to the coalescent. The model uses the NY98 codon model of molecular evolution which incorporates the ratio of non-synonymous to synonymous substitution, dN/dS. Values of dN/dS less than one are interpreted as purifying selection imposed by functional constraint and values greater than one are interpreted as diversifying selection imposed by interaction with the host immune system. Those sites under strong diversifying selection are predicted to be the major determinants of immunogenicity for the gene product. In order to exploit information about the underlying tree structure and mode of selection at sites segregating for insertions/deletions, I have described a simple extension to the NY98 mutation model. I have proposed a model for the variation in the dN/dS ratio and recombination rate along a sequence and a reversible-jump MCMC scheme for exploring that variation. The primary aim of the Bayesian inference framework described is to obtain a posterior distribution for the dN/dS ratio and recombination rate for every site along the sequence, but the underlying mutation rate, transition:transversion ratio and rate of insertion/deletion are also estimated. Finally I performed simulation studies to assess the performance of the inference method for two caricatures of variation in the dN/dS ratio and recombination rate. The method was found to have good coverage for the dN/dS ratio, but some upward bias in estimates of the recombination rate, in agreement with previous work. Most importantly, the simulation studies showed that when there is variation in the dN/dS ratio or recombination rate it can be detected, when there is no variation none is

detected, and there is little or no confounding between dN/dS and the recombination rate.

In the next chapter I will apply the new method to the *porB* locus of *N. meningitidis*, which encodes the antigenic PorB outer membrane protein. I will give a brief background to *porB* and the results of previous phylogenetic estimates of variation in the dN/dS ratio at the locus. In order to verify the conclusions of the *porB* analysis with the new method, I will apply a variety of model criticism techniques including prior sensitivity analysis and goodness-of-fit testing. Goodness-of-fit testing requires datasets to be simulated under the new model, so I will describe how to do that. I will briefly investigate the effect of violating the coalescent assumption of random sampling by comparing datasets of *porB* that represent a random and non-random sample. Finally, I will compare the results of the new method to previous phylogenetic methods to look for evidence of false positives caused by the assumption of no recombination.