

## Chapter 5

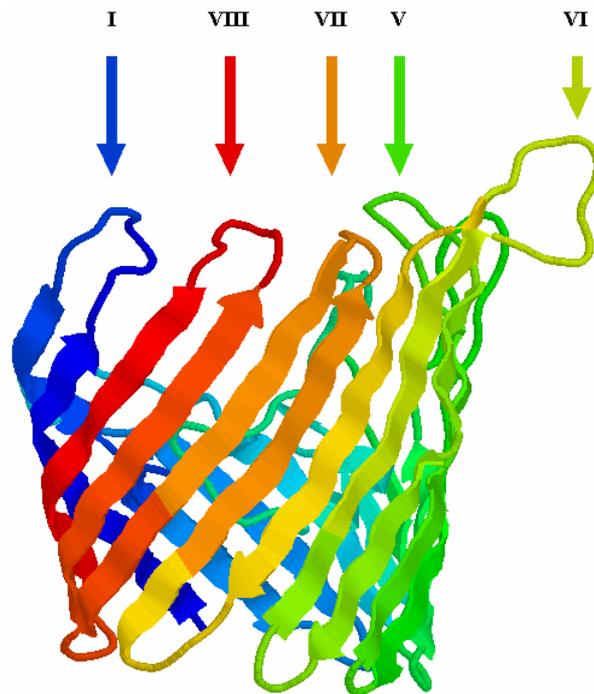
### Evidence for Immune Selection in an Antigen of

#### *Neisseria meningitidis*

Using the *porB* locus of *Neisseria meningitidis*, this chapter demonstrates the application of the Bayesian approach to inferring selection and recombination introduced in Chapter 4. PorB is ideal for exploring the new method, because it is a strongly immunogenic, constitutively expressed outer membrane protein whose molecular structure has been elucidated. Previous studies have analysed *porB* and I will compare the results of my method to those published before. In particular, I will compare the results to those of phylogenetic-based analyses for evidence of false positives introduced by the assumption of no recombination. Using *porB* as an example, I will outline a coherent approach to model-based analysis, from rejection of a model with no recombination through to prior sensitivity analysis and model criticism. Using different datasets there is the opportunity to informally study the effect of violating the coalescent assumption of random sampling. I will also contrast the patterns of variation in selection pressure in *porB* to those in the seven MLST housekeeping loci.

#### 5.1 Analysis of the *porB* locus

PorB is a porin expressed on the surface of the meningococcus, and thought to be important both for proper cell growth and pathogenesis. There exist two classes of



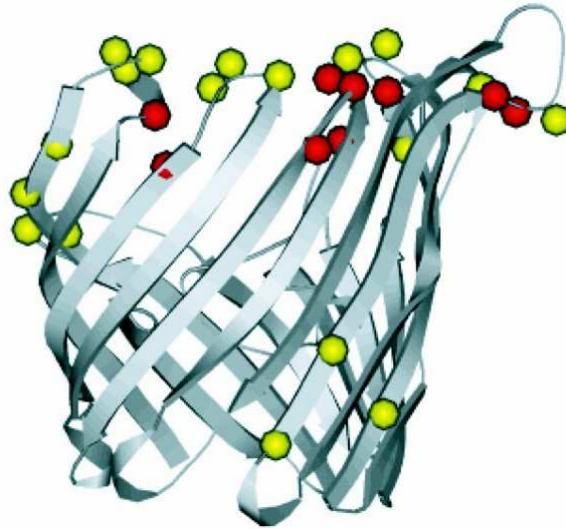
**Figure 1** Molecular structure of the *Neisseria meningitidis* class 3 outer membrane protein, PorB3. The molecule is colour-coded from the N (blue) to the C (red) terminus. The molecule spans the outer membrane, with eight exposed variable loop regions, of which five are marked (I, V, VI, VII and VIII). This image was generated using Protein Explorer (Martz 2002). The molecular structure was determined by Derrick *et al.* (1999). Jeremy Derrick kindly provided the molecular structure file.

PorB protein with somewhat different molecular structure and evolutionary ancestry (Smith *et al.* 1995; Derrick *et al.* 1999), called PorB class 2 and PorB class 3, or PorB2 and PorB3. These classes are defined on the basis of sequence homology and immunological properties (Smith *et al.* 1995). PorB is an important outer membrane protein (OMP) which is expressed constitutively at high levels, and is strongly immunogenic; epitopes of PorB define the serotypes of *N. meningitidis*. The molecular structure of meningococcal PorB comprises eight highly variable surface-exposed loop regions (I–VIII), consisting mainly of hydrophilic amino acid residues, between nine highly conserved membrane-spanning  $\beta$ -sheets. Figure 1 shows the

molecular structure of the PorB3 molecule (Derrick *et al.* 1995). The molecule is oriented with the surface-exposed regions at the top, and is colour-coded from the N (blue) to the C (red) terminus. Five of the eight loop regions are indicated by arrows.

### 5.1.1 Previous analyses

PorB is encoded by the *porB* locus. There have been several studies into the influence of natural selection on the genetic diversity of *porB2* and *porB3* alleles (Smith *et al.* 1995; Urwin *et al.* 2002). In an analysis based on 5 *porB2* sequences and 4 *porB3* sequences, Smith *et al.* (1995) counted the number of synonymous and non-synonymous differences to estimate that the relative rate of synonymous and non-synonymous change was  $dN/dS = 1.2$  for *porB2* and  $dN/dS = 0.62$  for *porB3* across the sequences as a whole. These values represent an average over the conserved and variable regions. Loop regions exhibited an elevated number of non-synonymous relative to synonymous substitutions; estimated to be 2.3 compared with 0.28 for non-loop regions, averaged over the *porB* classes. As discussed in Chapter 4, an elevated rate of non-synonymous change is indicative of relaxed functional constraint, or positive selection for variation in the amino acid sequence (diversifying selection). Evidence for diversifying selection in the surface-exposed loop regions is consistent with the hypothesis that the immune system exerts a selection pressure for antigenic novelty in PorB. However, on the basis of these estimates, that selection pressure is not enormous, which would suggest that PorB may be a less important vaccine constituent than other OMPs, for example PorA (Urwin *et al.* 2002).



**Figure 2** Location of sites under weak (yellow circles) and strong (red circles) positive selection in the PorB3 molecule, inferred using CODEML. All but 3 positively selected sites lie in the surface exposed loop regions, and seven out of the eight loops contain some sites under positive selection. Cf. Figure 1. Source: Urwin *et al.* (2002).

Urwin *et al.* (2002) used a maximum-likelihood method (Yang *et al.* 2000; described in section 4.1.2.1) implemented in the CODEML program of the PAML package (Yang 1997) to infer selection in the *porB* locus, taking the *porB2* and *porB3* allelic classes separately. The aim of the analysis was to improve the estimates of the dN/dS ratio by using many more sequences, applying a likelihood-based model for the ancestry of the sequences, and obtaining an estimate for each site along the amino acid sequence. They found evidence for extremely high selection pressures in *porB* surface exposed loops, as great as in HIV-1 surface glycoproteins, whilst the membrane spanning regions were under strong purifying selection. Based on the model of variation in selection pressure described in section 4.1.2.1, Urwin *et al.* (2002) estimated that in *porB2*, 94% of sites were under purifying selection (dN/dS = 0.067), 4.5% of sites were under weak positive selection (dN/dS = 4.2) and 1.1% of sites were under strong positive selection (dN/dS = 18.6). In *porB3*, 95% of

sites were under purifying selection ( $dN/dS = 0.033$ ), 4.1% of sites were under weak positive selection ( $dN/dS = 3.2$ ) and 0.7% of sites were under strong positive selection ( $dN/dS = 13.9$ ). The likelihood ratio tests for positive selection were very highly significant for both *porB* classes. Figure 2 shows the location of sites identified as experiencing weak (yellow circles) or strong (red circles) positive selection in *porB3*. Almost all positively selected sites lie in the surface exposed loop regions, except for one weakly selected site between loops IV and V, and two between loops V and VI. There is evidence for some positive selection in all loops except III.

Urwin *et al.* (2002) also conducted an analysis of the frequency of recombination in *porB*. The method is similar to that of Holmes *et al.* (1999), described in Chapter 1. A maximum likelihood tree was estimated for each half of the *porB2* sequence. Phylogenetic incongruence was quantified as the difference in log likelihood,  $\delta$ , between the maximum likelihood (ML) tree for the first half and the ML tree for the second half fit to the first (tree topology estimated from second half of the sequence and branch lengths estimated from the first half of the sequence conditional on the topology). Two hypotheses were tested: Firstly, the hypothesis that there is no recombination so that the topology is the same in both halves of the sequence. Secondly, the hypothesis that there is so much recombination that the topology estimated for the second half of the sequence is no better a fit to the first half of the sequence than a random topology. Both hypotheses were rejected for *porB2*, and for *porB3*, indicating that there is frequent recombination but not to the extent that the phylogenetic signal is utterly obliterated. Recombination in *porB* creates multiple, correlated, evolutionary histories for different parts of the sequence. The problem of this for CODEML is that it can inflate the false positive rate for detecting positive

selection (Anisimova *et al.* 2003; Shriner *et al.* 2003). The aim of the new method presented in Chapter 4, and implemented in the program omegaMap, is to co-estimate the dN/dS ratio and recombination rate along a sequence, with a more flexible model of variation in these parameters. In the sections that follow I apply the new method to the *porB3* sequences in order to visualise the variation in dN/dS and recombination rates, and demonstrate a coherent approach to testing the fit of the model so that it might be falsified. In the process I will compare the results to those of the CODEML analysis to look for potential false positives.

### **5.1.2 Isolates**

For the analysis I used the 79 *porB3* alleles sequenced by Urwin *et al.* (2002). The 79 alleles were sequenced from an assorted collection of isolates including carriage and disease from around the world. As a result, the 79 alleles do not constitute a random sample of any population in a meaningful sense, thus violating one of the assumptions of the coalescent model. The effect of violating the coalescent assumption of random sampling is unknown. Therefore only a subset of the 79 alleles was used: 37 alleles sequenced from isolates obtained during a swabbing programme at a military recruit training camp. Nasopharyngeal swabs were taken from healthy recruits several weeks after arriving at the camp. The catchment area of the training camp was England and Wales.

Of the 37 isolates, 19 were obtained by repeatedly swabbing 5 of the carriers; the remaining 18 were sampled from one carrier each. In Chapter 1 a metapopulation model was described in which each host corresponded to a single deme. It was shown that when no more than one isolate is sampled from each host, the ancestry of the

**Table 1 Permutation test for recombination**

	Carriage study		Global study	
	Correlation	$p$	Correlation	$p$
$r^2$	-0.18	0.001	-0.15	0.001
$D'$	-0.24	0.001	-0.16	0.001
$G4$	-0.23	0.001	-0.15	0.001

isolates can be modelled as a coalescent process in which the effective population size is a function of several epidemiological parameters including the duration of infection and the primary and secondary infection rates. Therefore, the collection of 37 isolates was thinned to 23 so that each host was represented by only one isolate each. In this chapter I will refer to this sample of 23 alleles as the *carriage study*, and to the full collection of 79 alleles as the *global study*. Whereas the global study consisted of 77 unique haplotypes, the carriage study consisted of 12 unique haplotypes. Rachel Urwin kindly provided her sequence alignments for the analyses presented in this section.

### 5.1.3 Test for recombination

Urwin *et al.* (2002) have already rejected the hypothesis that there is no recombination in the *porB3* alleles by quantifying phylogenetic incongruence within the sequences. The permutation test described in section 4.4.1 agrees with this conclusion for both the carriage and global studies. Table 1 shows the results. For the carriage study there was a 0.1% probability of observing as extreme a correlation

between physical distance and LD under the model of no recombination, regardless of choice of LD statistic, and the result was the same for the global study. Although the two methods reach the same conclusion, the permutation test for exchangeability of sites is much faster than the phylogenetic incongruence test because it does not involve estimating tree topologies.

#### **5.1.4 Codon frequencies**

In the inference scheme presented in Chapter 4, the codon equilibrium frequencies are not treated as parameters but rather as known constants. Specifying the equilibrium frequencies of the 61 non-stop codons can be done in several ways. One can make the simplifying assumption that all codons have equal equilibrium frequency. This is not supported empirically, but the assumption had little effect on the analyses presented in this chapter (data not shown). One can take the observed codon usage in the data. However, for a small dataset there will be considerable sampling error. This is not recommended because in the worst case, a codon may not be observed in the sample. Then adopting this approach would be to exclude the possibility of the codon appearing anywhere in the ancestral history of the sample, which is very undesirable. A possible solution is to take the observed nucleotide frequencies, and assume that codon frequencies are proportional to the product of their constituent nucleotide frequencies, but this is not supported empirically. In the analyses that follow the observed codon frequencies (Nakamura *et al.* 2000) in the complete genomic sequence of *N. meningitidis* Z2491 serogroup A (Parkhill *et al.* 2000) were used, after removing the stop codons. There are 730,000 codons in the meningococcal genome, which is more than adequate to overcome sampling problems.

**Table 2 Prior distributions**

	Prior A	Prior B
$\mu$	Exponential mean 0.07	Uniform 0 to 10
$\kappa$	Exponential mean 3.0	Exponential ratio
$\varphi$	Exponential mean 0.1	Exponential mean 1.0
$\omega$	Exponential mean 1.0	Gamma shape 2, scale 0.5
$\rho$	Exponential mean 0.1	Uniform 0 to 10

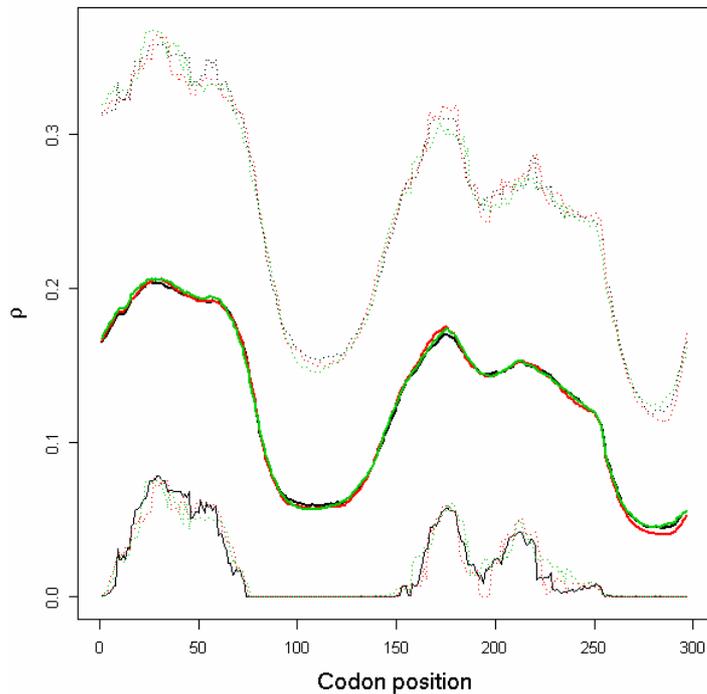
### 5.1.5 Priors

I chose to use exponential distributions for the priors on  $\mu$ ,  $\kappa$ ,  $\varphi$ ,  $\omega$  and  $\rho$ . Because the robustness of the results to the prior will later be tested by proposing an alternative prior, the prior specified here will be referred to as Prior A (Table 2). All the model parameters can take on any positive value. However, it might be more natural to consider them on a logarithmic scale. For example, the strength of purifying selection when  $\omega = 0.5$  might be interpreted as of an equal magnitude to the strength of positive selection when  $\omega = 2$ . This is because  $\omega$  is the ratio of the rate of non-synonymous mutation to synonymous mutation. Under the null model of selective neutrality, the rate of non-synonymous and synonymous mutation might be thought of as i.i.d. random variables, and as a result  $\omega$  or  $\omega^{-1}$  are equally valid parameterizations. I chose

a mean of 1 for the prior on  $\omega$  to represent the null model of selective neutrality. On a log scale, using the exponential distribution gives an approximately symmetric distribution. In fact the right tail decreases more rapidly than the left, so the prior favours less extreme values of  $\omega$  in the direction of diversifying selection than in the direction of purifying selection.

Because of the natural multiplicative interpretation of the other model parameters, and because for practical reasons it was originally simpler to program only a single distributional form for the priors, exponential distributions were fit to the other parameters. Other distributions were implemented for prior sensitivity analysis. It should be noted that for  $\omega$  and  $\rho$ , the reversible-jump MCMC scheme means that improper priors cannot be used because doing so forces the number of blocks to be 1 or  $L$ , ( $L - 1$  in the case of  $\rho$ ). The means of the priors were chosen based on analyses of *N. meningitidis* housekeeping loci (see Chapter 2). The mean of the prior on  $\mu$  was 0.07, and the mean for  $\kappa$  was put at 3.0. The rate of insertion/deletion was given a mean of  $\varphi = 0.1$ . For  $\rho$ , the mean was set at 0.1.

The prior on the number of blocks for  $\omega$  and  $\rho$  has already been described in section 4.2.5. The model for the variation in  $\omega$  and  $\rho$  can be considered a prior if the block structure is considered to be a parameter. That is to say that  $B_\omega$ ,  $B_\rho$ ,  $\mathbf{s}^{(B_\omega)}$  and  $\mathbf{s}^{(B_\rho)}$  are parameters and  $p_\omega$  and  $p_\rho$  would be hyperparameters of the priors on  $B_\omega$  and  $B_\rho$ . (See Chapter 4 Tables 1 and 2 for definitions of the notation.) If the block structure is considered to be missing data, then it is not a prior but a random effects model. The two are essentially equivalent in a Bayesian analysis. In order to specify that the average length of a block would be 10% of the sequence length ( $L = 298$  codons), the

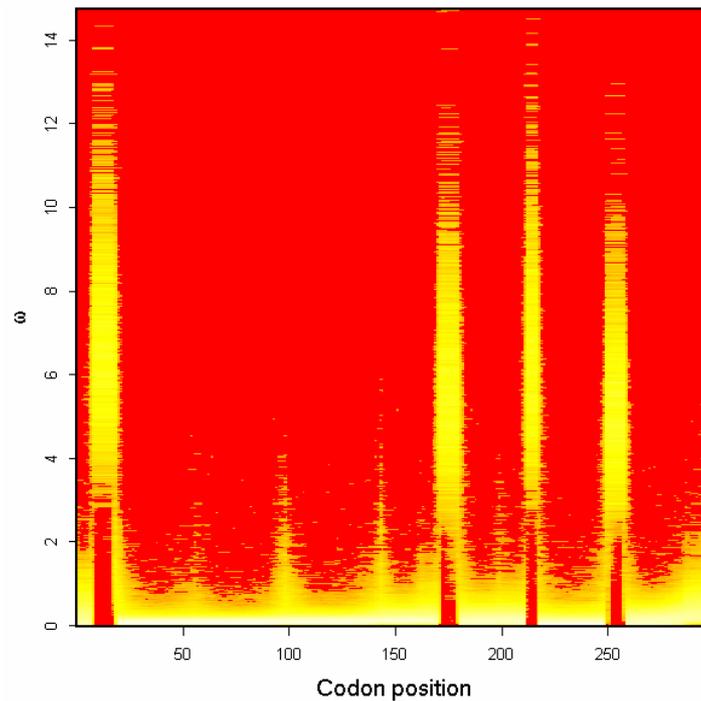


**Figure 3** Convergence of the marginal posterior for  $\rho$  along the sequence for the three MCMC chains (red, green and black) after 500,000 iterations. The burn-in was 20,000 iterations. For each codon the marginal mean and upper and lower 95% HPD bounds are shown.

prior on the number of  $\omega$  blocks was binomial with  $p_\omega = 1/30$ . Similarly, the prior on the number of  $\rho$  blocks was binomial with  $p_\rho = 1/30$ .

### 5.1.6 Results

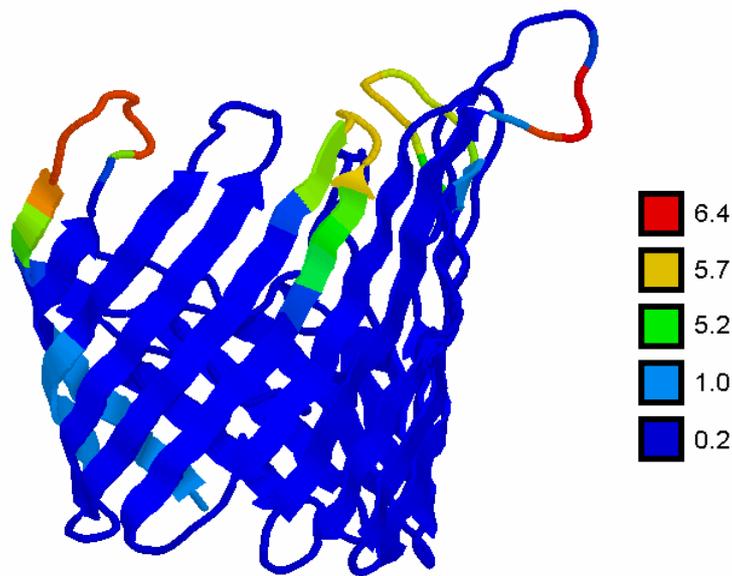
Three MCMC chains were run each for 500,000 iterations. Convergence was judged informally by comparing the posteriors obtained from the three chains. Figure 3 illustrates the posterior distribution of  $\rho$  along the sequence for the three chains, which are colour-coded red, green and black (the mean and 95% HPD interval is shown for each codon). A burn-in of 20,000 iterations has been removed from the start of each



**Figure 4** Fire-plot showing the site-wise posterior of  $\omega$  in the *Neisseria meningitidis* carriage study. More intense colours (closer to white) represent high posterior density, and less intense (closer to red) low posterior density.

chain. Figure 3 shows that the posteriors are in close agreement. Convergence is always best for the mean compared to the HPD bounds. Once convergence has satisfactorily been established for all the parameters, the chains are merged (minus the burn-in), and the results of the merged chains are presented.

Figure 4 shows a fire-plot for the posterior distribution of  $\omega$  at each site. More intense colours (closer to white) represent high posterior density, and less intense (closer to red) low posterior density. The structure of PorB3 consists of eight loop regions that extend out of the cell. Of these, there is clear and strong evidence for diversifying selection at four of the eight loops. In these loop regions the 95% HPD interval for the peak  $\omega$  is (3.58, 9.76), (3.01, 8.92), (3.26, 9.68) and (2.58, 7.57) for loops 1, 5, 6 and 7 respectively. Taking the point estimate of  $\omega$  at a site,  $\hat{\omega}$ , as the mean of the



**Figure 5** Molecular structure of PorB3 colour-coded according to  $\hat{\omega}$ , the mean of the posterior of  $\omega$  at each site. Dark blue indicates strong functional constraint and red indicates strong diversifying selection. This image was produced using Protein Explorer (Martz 2002).

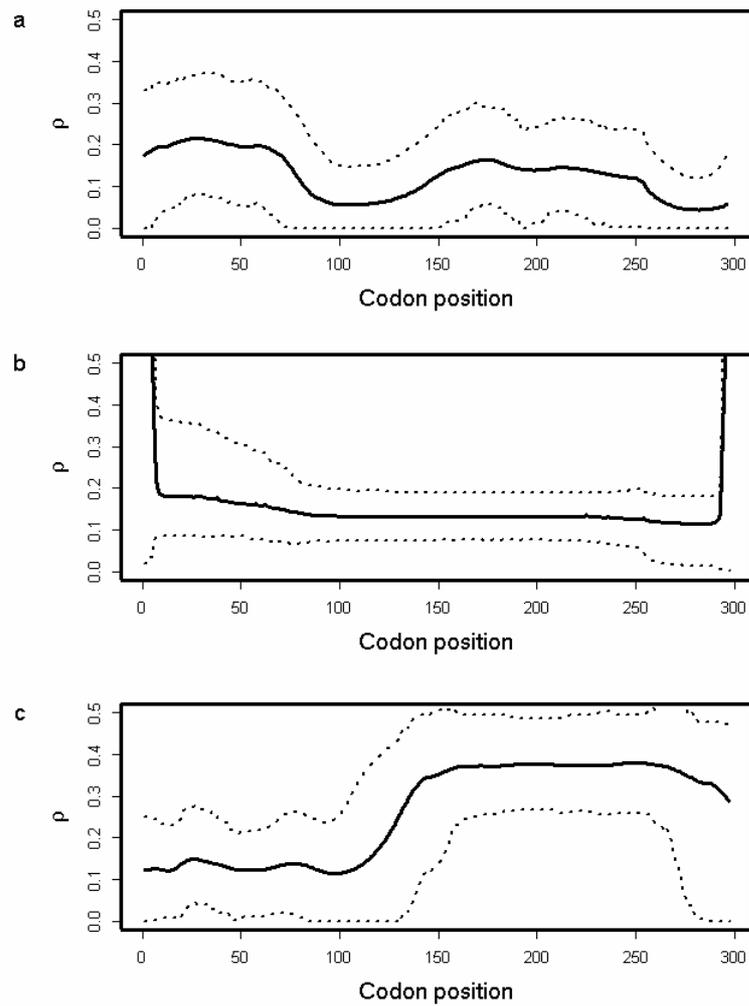
posterior distribution, then the average  $\hat{\omega}$  for the sequence is 0.90. Excluding sites for which  $\hat{\omega} > 1$ , this drops to 0.16. So the majority of the sequence is under strong functional constraint, but four of the eight loop regions are under strong diversifying selection.

Superimposing  $\hat{\omega}$  onto the three-dimensional structure of the PorB3 protein (Figure 5) illustrates the external position of loops 1, 5, 6 and 7. Because PorB3 is an outer membrane protein, these loops are especially exposed to the immune system, and are prime sites for recognition by antibody. It is striking that there is no evidence for diversifying selection outside the loops. Loops 2, 3 and 4 do not appear to be under diversifying selection; the three-dimensional structure suggests that they may be less exposed than the other loops. However, loop 8 is surprising because despite its prominent position (the dark blue loop second from left in Figure 5, cf. Figure 1),

**Table 3 Posterior distributions**

		Carriage study			Global study
		Prior A	Prior B	Prior A $\rho = 0$	Prior A
$\mu$	mean	0.27	0.35	0.45	0.31
	95% HPD	(0.18, 0.36)	(0.23, 0.48)	(0.33, 0.58)	(0.22, 0.40)
$\kappa$	mean	3.61	3.09	3.69	3.34
	95% HPD	(2.38, 5.00)	(1.94, 4.24)	(2.69, 4.83)	(2.41, 4.33)
$\varphi$	mean	0.09	0.17	0.29	0.08
	95% HPD	(0.02, 0.19)	(0.03, 0.37)	(0.08, 0.56)	(0.02, 0.16)
$R$	mean	37.7	46.8	-	78.0
	95% HPD	(27.2, 49.0)	(26.2, 75.0)	-	(61.6, 94.5)

there is very little support for diversifying selection between codons 280-295 (Figure 4). The light blue shading in Figure 5 occurs at the N and C termini, outside the nucleotide alignment I analysed. Therefore they have been assigned the mean of the prior,  $\hat{\omega} = 1$ .



**Figure 6** Posterior distribution of  $\rho$  in the carriage and global studies. The mean (solid line) and 95% HPD interval (dotted lines) are shown for (a) the carriage study under prior A, (b) the carriage study under prior B and (c) the global study under prior A.

There was some evidence for variation in the recombination rate (Figure 6a). The posterior mean for the total recombination distance,  $\hat{R} = 37.7$  (Table 3), was twice the prior mean of 19.9. The posterior on  $\mu$  was very different to the prior ( $\hat{\mu} = 0.27$ ), while there was little discrepancy for  $\kappa$  and  $\varphi$  ( $\hat{\kappa} = 3.61$ ,  $\hat{\varphi} = 0.09$ ).

## 5.2 Model criticism

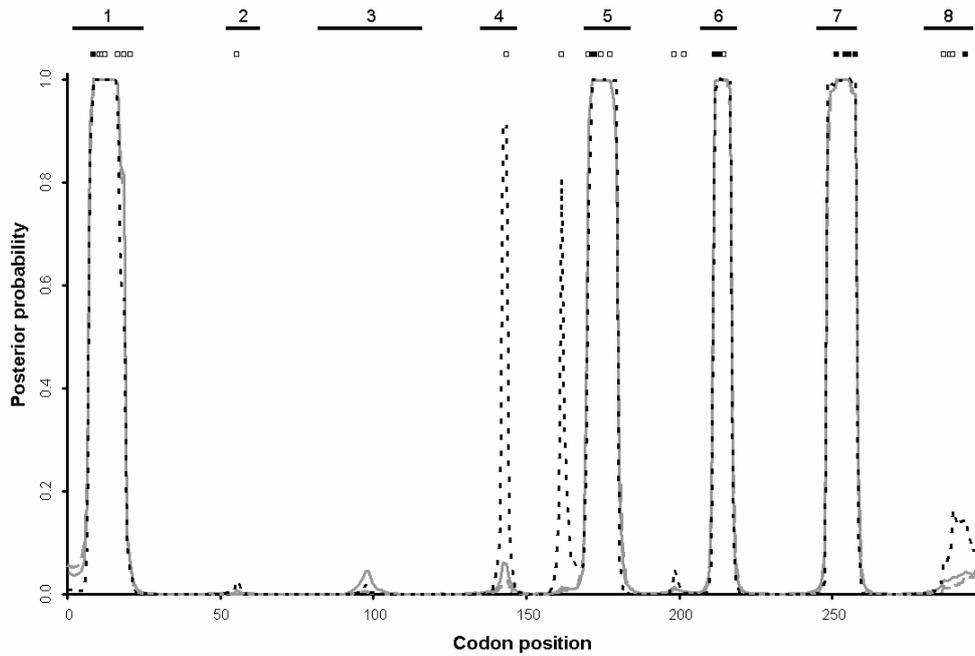
The application of phylogenetics-based techniques to detecting selection in genes sampled from within populations of microparasites such as viruses (Twiddy *et al.* 2002; Moury *et al.* 2004; de Oliveira *et al.* 2004) and bacteria (Peek *et al.* 2001; Urwin *et al.* 2002) has become widespread. However, many of these organisms are highly recombinogenic (McVean *et al.* 2002; Awadalla 2003), so the use of phylogenetic methods is inappropriate because a bifurcating tree is not an adequate model of the evolutionary history of these organisms (Holmes *et al.* 1999). This model misspecification can elevate the rate of false positives when detecting sites under positive selection (Anisimova *et al.* 2003; Shriner *et al.* 2003), which casts doubt over the validity of the inferences made. Model criticism is an important part of formulating and fitting an evolutionary model, because it allows that model to be falsified if it is a poor description of the data. In this section I discuss how model criticism can be integrated into the Bayesian analysis itself, using *porB3* as an example. The aspects of model criticism that I will focus on are the sensitivity of inference to the priors, goodness-of-fit testing using posterior predictive distributions of summary statistics, and the effect of violating the assumption of random sampling in the coalescent model.

### 5.2.1 Prior sensitivity analysis

To determine the sensitivity of inference to the choice of priors, the analysis was repeated with alternative priors (Prior B in Table 2). The choice of distributional forms and parameters for prior B is largely arbitrary, and necessarily so, because prior A was supposed to represent earnest prior beliefs. For practical purposes, a variety of

distributional forms was used to test the computer program. For  $\mu$  and  $\rho$  a uniform prior between 0 and 10 was fit (10 being the highest value considered plausible for either parameter). Following Huelsenbeck and Dyer (2004) a prior distribution on  $\kappa$  was fit that describes the ratio of two independent and identically distributed exponential random variables (see section 4.1.2.2). The moments, including the mean, for this distribution are undefined, but the median equals 1. For  $\varphi$  the mean of the exponential prior was changed from 0.1 to 1. Finally, for  $\omega$  a gamma distribution was used, still with a mean of 1, but with shape parameter 2, giving the distribution a mode at 0.5. This distribution retains the case of selective neutrality for its mean, but it tails off towards zero rather than increasing. Three MCMC chains were run, each 250,000 iterations in length, with a burn-in of 20,000 iterations. Having checked for convergence, the chains were merged to obtain the posteriors.

95% HPD intervals for the peak  $\omega$  in loops 1, 5, 6 and 7 show that the magnitude of the estimates has been reduced by the gamma prior to (2.76, 6.80), (2.16, 5.79), (2.31, 6.70) and (2.16, 5.66) respectively. Despite this, the relative height of the peaks is conserved. The average  $\hat{\omega}$  for the sequence is 0.68, reflecting the more conservative effect of the gamma prior. Excluding sites for which  $\hat{\omega} > 1$ , this drops to 0.17, which is almost identical to the inference based on prior A. This suggests that information about the absolute magnitude of sites under functional constraint is less influenced by the prior. Despite differences concerning the magnitude of  $\omega$ , the priors strongly agree on which sites are under diversifying selection (Figure 7).



**Figure 7** Site-wise posterior probability of diversifying selection ( $\omega > 1$ ) for the *porB3* carriage study, under prior A (grey solid line), prior B (grey dashed line), and prior A with the recombination rate forced to equal zero (black dotted line). The loop regions are numbered above. Those sites identified as under weak (empty squares) or strong (filled squares) positive selection by Urwin *et al.* (2002) are shown.

The posterior probability of diversifying selection at a given site is

$$\Pr(\omega > 1) = \int_1^{\infty} \Pr(\omega | \mathbf{H}) d\omega.$$

Prior A is represented by the solid grey line, and prior B by the dashed grey line. The two lines are virtually indistinguishable from one another at every site, indicating that identification of sites under diversifying selection is robust to the choice of prior.

Figures 6a and 6b compare the posterior probability of  $\rho$  given priors A and B. Under prior B, the posterior on  $\rho$  is somewhat flatter, with tighter credible intervals. The

average  $\hat{\rho}$  is largely the same for most of the sequence, except at the far ends, where  $\hat{\rho}$  increases sharply. This is an edge effect where, in the lack of information about the recombination rate, the posterior has been overwhelmed by the prior. The uniform prior on  $\rho$  has mean 5, explaining the rapid increase. The effect is reflected in the posterior on  $R$  (Table 3), which has a similar lower bound, but much increased upper bound. This striking sensitivity to the prior at the edges suggests that one should be cautious in interpreting the recombination rates at the edges of the sequence.

The posterior on  $\mu$  is influenced by the high mean of the uniform prior (Table 3), to the extent that  $\hat{\mu} = 0.35$  under prior B, which is only just inside the upper bound of the credible interval under prior A. In contrast,  $\kappa$  is not particularly sensitive to the prior, with largely overlapping credible intervals.  $\varphi$  shows a similar sensitivity to  $\mu$  in responding to a considerable increase in the prior mean. The lower bound is almost unaffected, but the mean and upper bound show a marked increase.

### 5.2.2 Posterior predictive $p$ -values

An essential property of any statistical model is that it should be falsifiable. A useful approach in Bayesian inference, and the one used here, is that of posterior predictive  $p$ -values (Rubin 1984). Here the model is taken to mean the probability model together with the posterior distribution of the model parameters. In essence, if the model is a good description of the data, then further datasets simulated under that model ought to resemble the real data. If they do not, then the model is failing in some important way. By *resemble* what is meant is that with respect to some statistic  $D$ , the

observed value of that statistic,  $D_{\mathbf{H}}$  should fall well within the range of values for the simulated datasets,  $D_{\mathbf{H}'}$ , where  $\mathbf{H}'$  is used to denote a simulated dataset.

The posterior predictive  $p$ -value is defined as the probability under the model of observing a discrepancy statistic  $D$  as large as that observed.

$$p = \int P(D_{\mathbf{H}'} \geq D_{\mathbf{H}} | \Theta)P(\Theta | \mathbf{H})d\Theta ,$$

where the integration is approximated by

$$p \approx \frac{1}{M} \sum_{i=1}^M I(D_{\mathbf{H}'_i} \geq D_{\mathbf{H}}). \quad (1)$$

In Equation 1,  $M$  is a large number (I used  $M \approx 15,000$ ),  $\mathbf{H}'_i$  is simulated from the posterior distribution  $P(\Theta | \mathbf{H})$ , and  $I$  is the indicator function. It is important to note that  $\mathbf{H}'_i$  is simulated under the exact probability model specified by the PAC likelihood and used in inference, which is not the coalescent but an approximation to it.

### 5.2.3 Simulating under a PAC model

The algorithm here for simulating under the PAC model follows directly from the description of the model (see section 4.2). Here the total number of sequences is denoted  $n$ .

1. Generate the first haplotype by drawing the codon independently at each site from the equilibrium frequency of codons. Let the number of haplotypes,  $k = 1$ .

2. Generate the  $(k + 1)$ th haplotype conditional on the first  $k$  by parsing the sequence 5' to 3' in the following way. For the 5'-most site, choose one of the  $k$  haplotypes to copy from uniformly at random. Call this haplotype  $x$ .
3. Mutate the current site  $j$  by drawing a time  $t$ , independently of all other sites, from an exponential distribution with rate  $k$ . Then use the transition probability matrix  $\mathbf{P}^{(t)}$  for the NY98 mutation model with parameters  $\mu$ ,  $\kappa$  and  $\omega_j$  to draw a codon conditional on codon  $j$  in haplotype  $x$  and time  $2t$ .
4. Move to the next site, and continue to copy haplotype  $x$  with probability  $(1 - \exp\{-\rho_j d_j / k\})$ , where  $\rho_j/2$  is the recombination rate between the sites (per bp per  $PN_e$  generations) and  $d_j$  is the distance (in bp). Otherwise choose one of the  $k$  haplotypes to copy from uniformly at random (including  $x$ ). Call this haplotype  $x$ .
5. Return to step 3 until at the 3'-most codon.
6. Let  $k = k + 1$ . If  $k < n$  return to step 2.

#### 5.2.4 Combining $p$ -values

A large number of datasets,  $M$ , are simulated as described in section 5.2.3. For each dataset the parameters are drawn from one of the iterations of the combined MCMC chain. Then for any particular discrepancy statistic, a marginal posterior predictive  $p$ -value can be calculated using Equation 1. The  $p$ -value is made two-tailed in the usual way. Combining  $p$ -values might be done using the Bonferroni correction, but this would be conservative even if the discrepancy statistics were independent. Since they are unlikely to be independent, Bonferroni would be too conservative for the goodness-of-fit test. That is to say that the  $p$ -value would under-estimate the extremity

of the observed data under the model, so the model would be less likely to be falsified. I am grateful to Jonathan Marchini for explaining the following way to combine  $p$ -values, which transforms the  $p$ -values into standard normal variates. By assuming that the transformed  $p$ -values can be made independent by removing the linear correlation structure (by further transformation of the multivariate normal distribution), a combined  $p$ -value can be obtained.

To combine two-tailed  $p$ -values for  $N$  different discrepancy statistics, denote the vector of discrepancy statistics for dataset  $j$

$$\mathbf{D}_j = (D_{1j}, D_{2j}, \dots, D_{Nj}).$$

Transform the marginal distribution of each discrepancy statistic  $i$  ( $D_{i1}, D_{i2}, \dots, D_{iM}$ ) into a standard normal distribution, so that

$$Z_{ij} = \Phi^{-1}\left(\frac{W_{ij} + 1}{M + 1}\right),$$

where  $W_{ij}$  is the marginal rank (with respect to  $j$ ) of discrepancy statistic  $D_{ij}$ , and  $\Phi^{-1}$  is the quantile function (inverse cdf) for the standard normal distribution. Next assume that the joint distribution of  $\mathbf{Z}_j = (Z_{1j}, Z_{2j}, \dots, Z_{Nj})$  is multivariate normal with zero mean and variance-covariance matrix  $\Sigma$ , where

$$\Sigma_{kl} = \begin{cases} r_{kl} & \text{if } k \neq l \\ 1 & \text{if } k = l \end{cases},$$

where  $r_{kl}$  is the correlation coefficient between the transformed discrepancy statistics  $k$  and  $l$  ( $Z_{kj}$  and  $Z_{lj}$ ) over datasets  $j$ . Next transform  $\mathbf{Z}_j$  to remove the correlation structure

$$\mathbf{Y}_j = \Lambda^{-1}\mathbf{Z}_j,$$

where  $\Lambda$  is obtained from the matrix factorization

$$\Sigma = \Lambda\Lambda^T.$$

$\Lambda$  is computed by singular value decomposition (see, for example, Press *et al.* 2002).

Include the observed values of the discrepancy statistics  $\mathbf{D}_H$  in the above procedure to obtain  $\mathbf{Y}_H$ . Assuming that the uncorrelated transformed discrepancy statistics are independent, then

$$X_j = \sum_{i=1}^N Y_{ij}^2$$

has a chi-squared distribution with  $N$  degrees of freedom. This can be verified by a histogram of the  $X_j$ 's. A one-tailed chi-square test of  $X_H$  combines the two-tailed posterior predictive  $p$ -values.

### 5.2.5 Choice of statistics and results

Discrepancy statistics have to be chosen that describe some aspect of the data that should be fit well by the model. This is important because it is unlikely that a model will fit all aspects of the data well. Statistics that are sensitive to mutation are  $S$ , the number of segregating sites and  $\bar{\pi}$ , the average number of pairwise differences. For recombination, the variance in the number of pairwise differences  $V(\pi)$  and the minimum number of recombination events  $R_m$  (Hudson and Kaplan 1985) are useful statistics (see section 2.1.3). The correlation between physical distance  $d$  and LD ( $r^2$ ,  $D'$  and  $G4$ ) that was used previously in the permutation test is also sensitive to recombination. For selection the statistic  $U$  is sensitive to any tendency for the

**Table 4** Posterior predictive  $p$ -values

	Carriage study				Global study	
	Observed	Prior A	Prior B	Prior A $\rho = 0$	Observed	Prior A
$S$	67	0.236	0.039	0.008	92	0.391
$\bar{\pi}$	25.3	0.340	0.179	0.003	26.9	0.068
$V(\pi)$	94.0	0.268	0.391	0.000	98.2	0.118
$R_m$	15	0.293	0.658	0.070	12	0.036
$\text{cor}(r^2, d)$	-0.13	0.247	0.265	0.002	-0.07	0.002
$\text{cor}(D', d)$	-0.24	0.440	0.353	0.000	-0.10	0.059
$\text{cor}(G4, d)$	0.22	0.443	0.332	0.000	0.09	0.144
$U$	0.5	0.543	0.878	0.711	0.5	0.621
$D$	1.05	0.121	0.058	0.567	0.97	0.398
Combined		0.268	0.103	0.001		0.013

simulated data to have too much or too little non-synonymous polymorphism on average.

$$U = \frac{\sum_{i=1}^L I(u_{\mathbf{H}'}^{(i)} \geq u_{\mathbf{H}}^{(i)})}{\sum_{i=1}^L I(u_{\mathbf{H}'}^{(i)} \neq u_{\mathbf{H}}^{(i)})},$$

where  $u^{(i)}$  is the number of non-synonymous pairwise differences minus the number of synonymous pairwise differences at site  $i$ .  $U$  should be centred around 0.5.  $U > 0.5$  indicates a bias towards diversifying selection and  $U < 0.5$  a bias towards functional constraint. Finally Tajima's  $D$  (Tajima 1989) is used, which is sensitive to directional

selection, balancing selection and demography; not forces that were modelled explicitly.

As with a classical  $p$ -value, if  $p$  is very small then the model does not fit the data well. Table 4 shows the observed values of all the discrepancy statistics and the two-tailed posterior predictive  $p$ -values for the carriage study under priors A and B. Of all the discrepancy statistics, the only posterior predictive  $p$ -value in the first two columns less than 0.05 is  $S$  for prior B. To obtain a single posterior predictive  $p$ -value for each model, the marginal  $p$ -values from one each of the mutation-sensitive, recombination-sensitive and selection-sensitive statistics ( $S$ ,  $\text{cor}(r^2, d)$  and  $U$ ) were combined following section 5.2.4. Table 4 shows that the combined posterior predictive  $p$ -values for the carriage study under priors A and B are  $p = 0.268$  and  $p = 0.103$  respectively. Neither is in the 5% tail of the distribution, suggesting the model fit is adequate with respect to mutation, recombination, and selection insofar as the dN/dS ratio is concerned. Tajima's  $D$  was positive ( $D = 1.05$ ), which may indicate balancing selection or population structure. The  $p$ -value for neither prior was in the 5% tail, so while these forces have not been modelled explicitly, the fit appears to be adequate. In fact for finite-sites mutation models, Tajima's  $D$  can have an expectation greater than zero under the standard neutral model

### **5.2.6 Analysis of the global study**

As an informal test of how violating the coalescent assumption of random sampling would affect inference, the 79-sequence PorB3 data (the global study) of Urwin *et al.* (2002) were analysed using prior A. For computational tractability one randomly

chosen ordering of the haplotypes was used. Three MCMC chains were run, each 500,000 iterations in length, with a burn-in of 20,000 iterations. Having checked for convergence, the chains were merged to obtain the posteriors. Table 4 shows that  $\hat{\mu} = 0.31$  was barely larger than for the carriage study, and the credible intervals overlapped almost entirely. The rate of insertion/deletion,  $\phi$  was not greatly affected ( $\hat{\phi} = 0.08$ ), nor was the transition-transversion ratio ( $\hat{\kappa} = 3.34$ ). But the total recombination rate doubled to  $\hat{R} = 78.0$  with no overlap in the credible intervals. Across the sites, the recombination map (Figure 6c) does not differ greatly in the left half of the sequence (c.f. Figure 6a), but thereafter rises rapidly to about  $\rho = 0.38$ . The low posterior predictive  $p$ -values for the recombination-sensitive discrepancy statistics (Table 4) advises caution on the interpretation of  $\hat{\rho}$ .

However, inference on  $\omega$  was hardly affected. Loops 1, 5, 6 and 7 still have very high posterior probabilities of diversifying selection. The magnitude of  $\omega$  inferred for each loop is comparable, with the 95% HPD intervals for the four loops (2.89, 7.28), (3.47, 8.17), (3.22, 8.79) and (3.10, 7.60). The only substantive difference is in loop 8, which now also has high posterior probability of  $\omega > 1$ . The 95% HPD interval for the peak  $\omega$  in loop 8 is (0.66, 2.87) and  $\Pr(\omega > 1) = 0.92$ . This difference can be explained by sites in loop 8 that exhibit amino acid variation in the global study but not the carriage study. The average  $\hat{\omega}$  for the whole sequence is 0.91, and excluding sites for which  $\hat{\omega} > 1$ , it drops to 0.22, both comparable to the carriage study.

### 5.3 Evidence for false positives

Ancestral recombination can cause false positives in phylogenetic methods (Shriner *et al.* 2003, Anisimova *et al.* 2003). If this has had an important effect on the analysis of meningococcal PorB3 then one should expect to see those false positives when the results of the CODEML analysis (Urwin *et al.* 2002) are compared to those presented here. Those sites identified as under weak (empty squares) and strong (filled squares) diversifying selection by CODEML are illustrated in Figure 7. All of the strongly selected sites and all but five of the weakly selected sites fall within loops 1, and 5-8. With the exception of loop 8 all these sites had high posterior probability of diversifying selection for the carriage study (Figure 7). When the global study is analysed, loop 8 also has high posterior probability of diversifying selection. Therefore there are just five sites where CODEML inferred diversifying selection but omegaMap did not. These are candidates for false positives.

There are a number of possible explanations for discrepancies of this kind, including

1. The approximation in omegaMap has given rise to false negatives. The PAC likelihood does not explicitly model the genealogy and this might have unexpected effects.
2. The block-like prior in omegaMap caused false negatives. Imposing a model in which adjacent sites share a common selection parameter might disfavour isolated sites under diversifying selection.
3. Recombination has caused CODEML to give false positives.

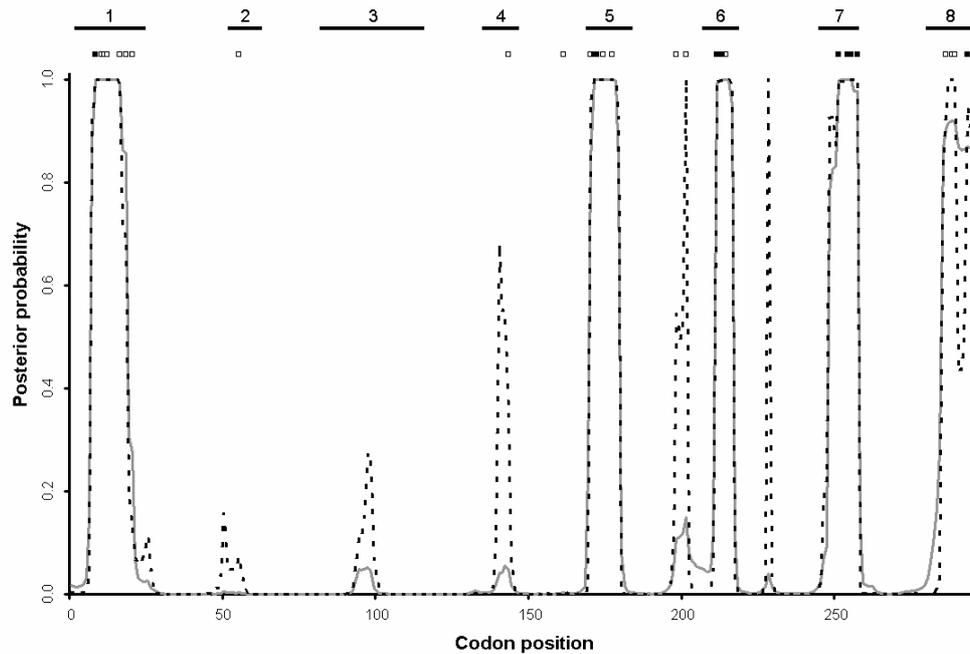
In an attempt to distinguish between the explanations, an analysis of the carriage study was performed in which the recombination rate was forced to equal zero. Using

prior A, three chains were run for 500,000 iterations each. After a burn-in of 20,000 iterations the chains were compared for convergence and merged to give the posterior.

In Figure 7 the site-wise posterior probability of diversifying selection is plotted (black dotted line) for comparison with the other analyses. The false-positive candidates are located at sites 55, 143, 161, 198 and 201. Of these, the first two are located in loops 2 and 4 respectively. The remaining three are not in loops. Comparison of Figure 4 to Figure 2 shows that these latter three disputed sites are located in a cytoplasmic region of the protein. The site-wise posterior probability of diversifying selection is very similar to the other analyses (Figure 7), except at two positions. These two positions correspond to two of the five false-positive candidates: sites 143 and 161. Although one cannot be certain that these sites are false positives, the results are suggestive.

The posterior predictive  $p$ -values (Table 4) show that the deleterious effect of assuming no recombination is not confined to recombination-sensitive discrepancy statistics. The mutation-sensitive parameters also have extremely low  $p$ -values (0.008 and 0.003 for  $S$  and  $\bar{\pi}$  respectively). The combined test shows that the model as a whole is a very poor description of the data ( $p = 0.001$ ). Although the selection-sensitive parameters do not have significant  $p$ -values, the consequence of the model inadequacy is to cast doubt on all inferences made from it.

The PAC model in the absence of recombination does not default to the coalescent with no recombination because the tree is still not modelled explicitly. Therefore it is unlikely that the assumption of no recombination will affect a PAC model and a



**Figure 8** Site-wise posterior probability of diversifying selection ( $\omega > 1$ ) for the *porB3* global study, under prior A (grey solid line), and prior A with the recombination rate forced to equal zero (black dotted line). The loop regions are numbered above. Those sites identified as under weak (empty squares) or strong (filled squares) positive selection by Urwin *et al.* (2002) are shown. Cf. Figure 7.

phylogenetic model in an exactly equivalent fashion. Nevertheless, when it is assumed that there is no recombination, sites that otherwise had low posterior probability of diversifying selection attained high posterior probabilities. This outcome is exactly what is predicted by the work of Shriner *et al.* (2003) and Anisimova *et al.* (2003).

When the global study is analysed under the constraint that the recombination rate equals zero, the same effect is seen: sites that have low posterior probability of diversifying selection when the recombination rate is unconstrained attain high posterior probabilities. But the actual sites affected differ when the global study is

analysed compared to the carriage study (compare dotted line in Figure 8 to Figure 7). Of the false positive candidates, site 55 remains at low posterior probability. Sites 143 and 161 have reduced posterior probabilities, 0.35 compared to 0.91, and 0.01 compared to 0.59, respectively. Sites 198 and 201 have increased posterior probabilities, 0.54 compared to 0.02, and 0.99 compared to 0.00, respectively. In addition, sites 140 and 228 reach high posterior probability. These sites exhibit amino acid polymorphism in the global study but not the carriage study. Loop 8 sites also show high posterior probabilities, compared to the carriage study with zero recombination. Like sites 140 and 228, these sites exhibit amino acid polymorphism in the global but not the carriage study, but are not candidates for false positives because they are identified as under positive selection in the global study with unconstrained recombination.

The effect on inference of assuming no recombination is complex, and as a result it is not possible to say with any confidence that particular sites are false positives. In general, however, by constraining the recombination rate to equal zero in the PAC models, the number of sites identified as under positive selection is inflated, which is predicted by the work of Anisimova *et al.* (2003) and Shriner *et al.* (2003). Only sites that exhibit amino acid polymorphism or immediately adjacent sites are identified as under positive selection. Why might imposing a phylogenetic tree on recombining genes lead to some sites that exhibit amino acid variation to have an inflated estimate of the dN/dS ratio? Recombination can cause homoplasies in a phylogenetic tree, i.e. a pair of sequences that appear to be distantly related overall share a rare allele at a particular site which the rest of the sample suggests was not shared by the common ancestor of those sequences. Therefore, recurrent mutation must be invoked to explain

the homoplasy. (See for example Figure 4 in Chapter 4; homoplasies can be explained by recombination or by extra mutation). As a result, some sites will appear to be hypermutable. If  $\omega$  is allowed to vary but the synonymous rate of mutation is not, then sites with non-synonymous homoplasies will be best fit by an elevated  $\omega$  at that site. These sites will appear to be positively selected. Yet sites with synonymous homoplasies, which will also appear hypermutable, cannot be fit by simply lowering  $\omega$  because the synonymous mutation rate is constrained. Therefore there is no symmetric effect in which some sites have very low  $\omega$ .

## 5.4 Analysis of housekeeping loci

Evolutionarily, antigen genes such as *porB* and housekeeping genes responsible for essential metabolic processes are under entirely different selection regimes. Whilst an antigen such as *porB* is exposed to immense selection pressure for antigenic novelty imposed by the host immune system, a housekeeping gene is, perhaps as a result, shielded from such co-evolutionary conflict and has the opportunity to adapt to an optimal functional form provided the necessary mutations arise. Therefore one would expect that housekeeping loci exhibit strong functional constraint. The loci used in MLST were chosen to be conserved, with no unusual signatures of selection or recombination, but sufficiently polymorphic to provide resolution for typing (Urwin and Maiden 2003). Contrasting the variation in the dN/dS ratio and recombination rate (selection and recombination ‘maps’, say) for the *porB* locus with those estimated from the MLST loci should make for interesting comparisons.

**Table 5 Genetic diversity in Czech carriage study**

Locus	Length (bp)	No. alleles	No. poly- morphic sites	dN/dS
<i>abcZ</i>	432	21	75	0.074
<i>adk</i>	465	19	25	0.011
<i>aroE</i>	489	21	135	0.295
<i>fumC</i>	465	29	48	0.010
<i>gdh</i>	501	19	26	0.049
<i>pdhC</i>	480	25	83	0.068
<i>pgm</i>	450	24	80	0.112

Table 5 summarises the genetic diversity in the seven MLST loci sequenced from a population of carried meningococci in the Czech Republic in 1993 (Jolley *et al.* 2000). There were 217 isolates in total. Polymorphism at the nucleotide level ranges from 5-20% across the loci. The dN/dS ratios, estimated using the number of non-synonymous and synonymous polymorphisms observed in the data, range from 0.010-0.295. *fumC* is the least polymorphic locus and has the lowest dN/dS ratio. *aroE* is the most polymorphic locus and has the highest dN/dS ratio. All dN/dS ratios are less than one, suggesting that differences may be due to the relative functional constraint rather than there being any evidence for positive selection. However, because these ratios are averaged across the sequence any positive selection occurring might be diluted by surrounding highly constrained regions.

For each locus, a random subset of fifty out of the 217 Czech sequences was chosen for analysis. This mainly reflects a computational constraint in that the complexity of

the PAC likelihood is quadratic in the number of sequences (Chapter 4). For the same reason, only one ordering of the PAC likelihood was used in these analyses. For direct comparison with the *porB3* results, the same priors on  $\mu$ ,  $\kappa$ ,  $\varphi$ ,  $\omega$ ,  $\rho$  (Prior A, Table 2) and the number of blocks were used. This is not entirely justifiable because the Czech carriage data were used originally to inform the choice of priors on *porB3*. For each locus three MCMC chains were run, each of length 500,000 iterations, with a burn-in of 20,000 iterations. Having established convergence, the three chains for each locus were merged to obtain the posteriors.

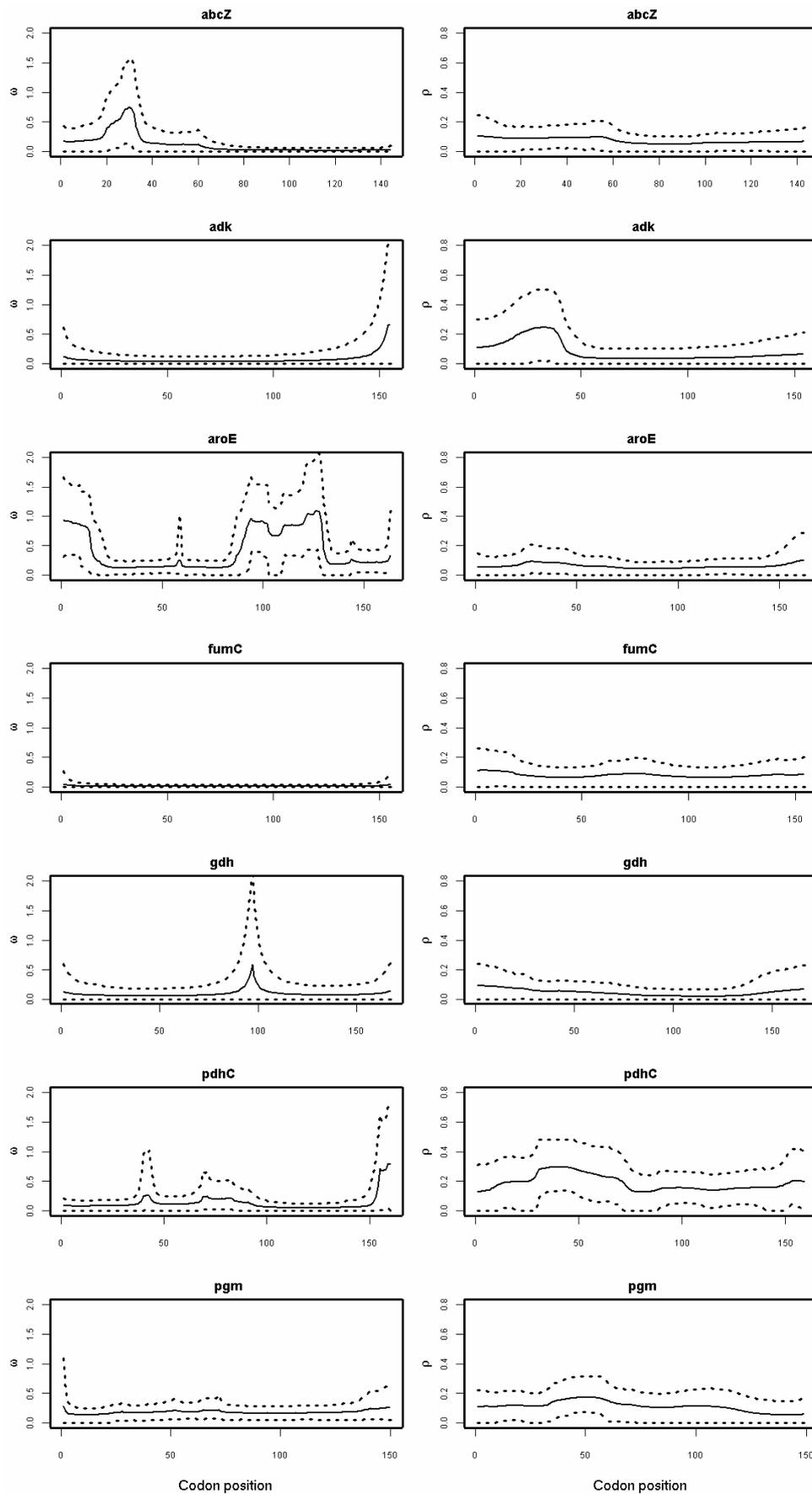
Figure 9 shows the estimated selection and recombination maps for the seven loci. From the selection maps it is clear that the sequence-averaged dN/dS ratios presented in Table 5 belie the true extent of variation in selection pressure in some of the housekeeping genes. Recombination rates are more conserved, with little compelling evidence for major peaks or troughs in any of the loci. There is some evidence for limited recombination rate variation in *adk* and *pdhC*. Interestingly, *fumC* is particularly functionally constrained, with almost no variation in  $\omega$  along the sequence (average  $\hat{\omega} = 0.016$ ) and *aroE* exhibits the greatest variation in and highest values of  $\omega$  (average  $\hat{\omega} = 0.409$ ). The *abcZ*, *gdh* and *pdhC* loci all show some spikes in the selection map.

**Table 6 Point estimates for MLST loci**

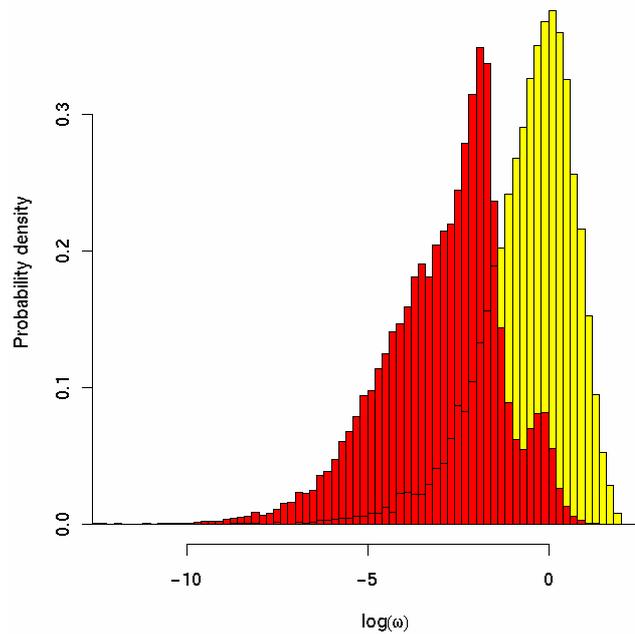
	<i>abcZ</i>	<i>adk</i>	<i>aroE</i>	<i>fumC</i>	<i>gdh</i>	<i>pdhC</i>	<i>pgm</i>
$\bar{\omega}$	0.124	0.076	0.409	0.016	0.101	0.136	0.184
$\mu$	0.298	0.106	0.696	0.195	0.113	0.452	0.601
$\kappa$	10.2	7.63	3.38	9.51	6.88	5.86	3.77
$\varphi$	0.056	0.050	0.257	0.071	0.097	0.051	0.159
$R$	10.7	12.6	10.0	12.3	7.86	29.9	16.7

Table 6 shows the point estimates (mean posterior) for  $\mu$ ,  $\kappa$ ,  $\varphi$ ,  $R$  and the average  $\omega$  ( $\bar{\omega}$ ) along the sequence for each locus. The estimates of  $\bar{\omega}$  agree well with the relative estimates of dN/dS in Table 5, but not the absolute values. Likewise, the estimate of  $\mu$  (Table 6) agrees well with the relative polymorphism in the seven loci (Table 5). The recombination rates are of the same order of magnitude as those estimated in Chapter 2, but appear to be slightly elevated, possibly a manifestation of the upwards bias noted in Chapter 4. The relative magnitude of recombination rates across loci agree well.

There is little evidence for diversifying selection in any of the MLST loci in the Czech carriage study (Figure 9), although sections of *aroE* appear to be close to selective neutrality. Some loci appear to show an increase in  $\omega$  at the 5' and 3' extremes of the sequence. This is most likely an edge effect, caused by reduced sequence information at the extremes so the prior has more influence over the inferred dN/dS ratio. In no locus does the mean of the posterior appear to exceed 1 at the extremes of the sequence.



**Figure 9** Variation in  $\omega$  and  $\rho$  in MLST genes in the Czech carriage study.



**Figure 10** Red histogram: posterior distribution of  $\log(\omega)$  amalgamated across MLST loci in the Czech carriage study. Yellow histogram: prior distribution of  $\log(\omega)$  for comparison. The prior on  $\omega$  is an exponential distribution with mean 1.

On the whole, the MLST loci are functionally constrained, which is not unexpected. Whilst the prior favoured selective neutrality, the posterior distributions clearly favour lower values of  $\omega$ . This is illustrated on a logarithmic scale in Figure 10. The posterior distribution of  $\log(\omega)$  has been amalgamated across sites and across loci (red histogram). For comparison, the exponential prior with mean 1 is plotted (yellow histogram). There are two patterns of note. The whole posterior distribution is left-shifted compared to the prior, indicating good evidence of purifying selection in these housekeeping genes. However, the amalgamated posterior is bimodal, with a small peak at  $\log(\omega)=0$ , suggesting that a minority of sites are actually selectively neutral, presumably due to relaxed functional constraint. There is no evidence for diversifying

selection in the amalgamated posterior. Figure 10 might serve as a useful prior in future studies of variation in  $\omega$  in housekeeping loci.

## 5.5 Summary

In Chapter 4 I presented a new method for estimating the selection parameter  $\omega$  and the recombination rate  $\rho$  from a sample of gene sequences. Uncertainty in the evolutionary history was taken into account using a coalescent-based approximate (PAC) likelihood. Variation in  $\omega$  and  $\rho$  was modelled as a block-like structure with a variable number of blocks. A Bayesian inference scheme was used to average over the number and position of the blocks using reversible-jump MCMC and obtain the posterior distribution of the other parameters. Using simulations, the new method showed good power to detect variation in  $\omega$  and  $\rho$ , and did not appear to confound the two. The method has a low false positive rate for detecting sites under diversifying selection. In this chapter I applied the method to the *porB* locus of *Neisseria meningitidis* and performed prior sensitivity analysis and model criticism to verify the results. I constrained the recombination rate to equal zero to discover the effect on inferring positive selection, and compared the results to those of phylogenetic methods that assume no recombination. The comparison suggests that some sites identified as under positive selection by CODEML may have been false positives. Seven housekeeping loci were analysed from a carriage study in the Czech Republic, and the inferred levels of selection and recombination were compared to those observed in the antigen gene *porB*. There was no evidence for diversifying selection in any of the housekeeping genes.