Chapter 6

Further Developments

Two distinct roles for mathematical modelling of meningococcal biology have been pursued in this thesis. On the understanding that any model is a caricature of reality, the pertinent question is not whether the model is right or wrong, but which aspects of reality does it fail to adequately capture? Answering this question is the first of those roles, and the one that I have utilised for learning about the population structure of *Neisseria meningitidis*. Beginning with a standard neutral coalescent model, which, I argued in Chapter 1, is the appropriate starting point for modelling infectious microorganisms, I used cross-validation to demonstrate that carried populations of meningococci have greater than expected population structuring. As seen in Chapter 2, that structure manifests itself as a dearth of unique sequence types (STs), positive values of Tajima's *D*, and stronger than expected correlation between linkage disequilibrium (LD) and physical distance. In Chapter 3 I showed that the excess structuring cannot be explained simply by over-sampling of closely related meningococci, as proposed by the neutral microepidemic model. Description of meningococcal populations using AMOVA and the Mantel test revealed that within Bavaria, there is some genetic differentiation caused by isolation by distance. By contrast, in the Czech Republic there was no evidence for genetic differentiation, whereas between the European countries of the Czech Republic, Greece and Norway, there is significant differentiation, the nature of which is not simple isolation by distance but reflects more complex sub-continental transmission routes. Disease-

causing meningococci do not appear to be a random subset of carried meningococci; observed genetic diversity suggest that disease-causing meningococci may persist alongside asymptomatic meningococci. In the discussion below I concentrate on how the results of these descriptive analyses can be used to refine a coalescent-based model of meningococcal populations.

Second of the two roles is estimating parameters of evolutionary relevance. When the model is a good fit, which was established in Chapter 5 using posterior predictive *p*-values, the parameter estimates can be biologically interpreted. Fitting the model of immune selection introduced in Chapter 4 revealed variation in selection pressure but not recombination rate within meningococcal antigen genes. Amino acid residues under strong diversifying selection corresponded to extracellular loop regions in the tertiary structure of the outer membrane protein PorB. This is consistent with selection for antigenic novelty in regions of the protein exposed to the immune system. In meningococcal housekeeping genes the dN/dS ratio was estimated to be consistently below the neutral expectation of one, concordant with the view that their gene products are not under diversifying selection. Quantitative inference allows evolutionary parameters from different genes to be compared; in addition to inferring that housekeeping genes are under strong functional constraint, it was shown that there is much less variation in the dN/dS ratio compared to the *porB* locus. I based the model of immune selection on the NY98 codon model of molecular evolution (Nielsen and Yang 1998), together with the PAC likelihood model of Li and Stephens (2003) and a piecewise constant model of variation in the dN/dS ratio and recombination rate (Green 1995; McVean *et al.* 2004). Below I discuss the advantages

and limitations of the new model, including the implications for vaccine research, and possible future extensions to it.

6.1 Meningococcal population structure

To begin with I will discuss the advantages of explicit evolutionary models over descriptive analyses such as AMOVA and the Mantel test. Descriptive analyses can help to direct the refinement of coalescent-based evolutionary models, and I will go on to describe how the results of the descriptive analyses performed in Chapter 3 might be used to inform development of a coalescent model with population structure. Then I will pursue the way in which inference might be performed using approximate Bayesian computation with conditional density estimation (ABC-CDE).

6.1.1 Advantages of explicit evolutionary models

Methods of analysis such as AMOVA, Mantel tests and logistic regression are useful ways to explore patterns of genetic diversity in meningococcal populations, and allow various informal scenarios to be explored. Such exploratory analyses can be used to help inform a more coherent approach to understanding the evolution of meningococci. None of the methods listed above make use of explicit evolutionary models. The advantage of an explicit evolutionary model is that parameters of interest can be estimated and the deficiencies of the model can be explored in a way that is readily interpretable. For example, in the analyses presented in Chapter 3 it was claimed that there is no evidence for geographic differentiation within the Czech Republic but there is in Bavaria, on the basis of whether the permutation test yielded a significant *p*-value or not. If an explicit evolutionary model with migration were fitted

to the data, estimates of transmission rates between geographic localities could be directly estimated and the results compared in a quantitative fashion. Such an approach might be more informative than the polarised result obtained from the permutation test that there either is or is not structure.

Population structure can be incorporated into the coalescent in several ways. When migration rates are high, so that the per generation probability of migration *m* is larger than $O(1/N)$ where *N* is the population size within a geographic deme, but the number of geographic demes *D* is finite,

$$M = \lim_{N\to\infty} PNm = \infty .$$

In this scenario, known as the strong migration limit, the genealogy of the population is that of a standard coalescent model with an altered timescale (see for example Nordborg 2003). Unlike the separation of timescales result obtained for the large *D* approximation (Wakeley 1998, 2001; Wakeley and Aliacar 20001), there is no scattering phase adjustment for the configuration of the sample amongst demes. In Chapter 2 a model of meningococcal evolution with a standard coalescent genealogy was rejected. Thus, the strong migration model of population structure can be rejected for meningococcal evolution, as can any other model that results in a standard coalescent.

When the migration rate is $O(1/N)$, so that

$$M = \lim_{N\to\infty} PNm$$

is finite, the resulting genealogy is known as the structured coalescent (Wilkinson-Herbots 1998). In the structured coalescent the number of geographic demes *D* is

finite, and the relative population size of all demes as well as all pairwise migration rates between demes, need to be specified. Therefore the structured coalescent has a large number of parameters, although various simplifications can be made, such as symmetric migration (Wright 1931, Maruyama 1970), one or two-dimensional stepping stone models (Kimura and Weiss 1964), or the large *D* approximation (Wakeley 2001). Another way of overcoming the problem of many parameters in the context of meningococcal evolution is discussed in the next section.

Using the structured coalescent would allow relevant evolutionary and epidemiological parameters to be estimated. For example, if population structure within Bavaria were to be modelled by treating each sampling location as a separate deme, then in principal the rates of migration between all pairs of towns could be estimated. These rates of migration correspond to transmission rates between demes, which is of epidemiological relevance for understanding how meningococci spread through human populations. If disease-causing and carried meningococci were to be treated as separate demes, then the migration rate could be interpreted as the rate of genetic exchange between the two groups, or the degree of genetic isolation. This again is of epidemiological relevance, for example estimating the rate of acquisition of antibiotic resistance in bacteria, or the speed at which capsule switching may occur in response to a specific vaccine (see section 1.1.3.2).

Using explicit evolutionary models allows for more meaningful hypothesis testing. For example, in Chapter 1 I discussed the problems with using classical methods such as the χ^2 goodness-of-fit test for rejecting a model of linkage equilibrium. The χ^2 test is inappropriate essentially because the wrong null distribution is obtained; genetic

drift can cause deviations from linkage equilibrium even for unlinked loci. Therefore the null hypothesis of linkage equilibrium might be rejected when it is true; the test is anti-conservation. By contrast, the permutation test used by AMOVA to detect population differentiation in the Czech Republic is conservative. This might explain why genetic differentiation was detected between sampling locations in Bavaria but not the Czech Republic even though the two regions are roughly the same size. In the context of using an explicit evolutionary model, more powerful analyses might be obtained by constructing a likelihood ratio test in a classical setting, or using Bayes factors or posterior predictive *p*-values in a Bayesian setting, (although it might be necessary to utilize summaries of the data for computational reasons). In this example the evolutionary model would be the structured coalescent.

6.1.2 Bayesian inference in the structured coalescent

The parameters of the structured coalescent could be estimated in the ABC-CDE (approximate Bayesian computation with conditional density estimation) setting described in Chapter 2. In the most general formulation of the structured coalescent, there is a migration rate for each pair of populations, or demes. Currently there are no 'full-likelihood' methods that estimate migration rates in the coalescent in the presence of recombination using the full sequence data. The problem obviously gets more difficult with increasing number of demes, but one approach that could be harnessed in the ABC-CDE framework would be to restrict the number of free migration parameters to one average rate.

The approach that I propose is to pre-determine the relative pairwise migration rates according to the population sizes of those demes (which could be measured from

census data for human towns), the geographic distance of those demes (which is straightforward to measure for towns), or some other geographic measure of connectivity, which might be as simple as the shortest road distance between towns. Several summary statistics exist that would be sensitive to the migration rate, including F_{ST} and the correlation between genetic and geographic distance. The various ways of defining the relative pairwise migration rates could quite easily be compared in the Bayesian framework using Bayes factors, cross-validation (discussed in section 2.3.4) or posterior predictive *p*-values (discussed in section 5.5.2). In my opinion, coalescent-based models, which can be fitted using methods such as ABC-CDE, offer a common thread which, in an iterative framework of model criticism and refinement, will be the best way to improve understanding of the evolution and epidemiology of meningococci.

6.2 Detecting selection in *Neisseria meningitidis*

In Chapter 4 I introduced a new model for detecting selection in genes of interest in *N. meningitidis*. The model was based on the NY98 codon model of molecular evolution (Nielsen and Yang 1998), together with the PAC likelihood model of Li and Stephens (2003) and a piecewise constant model of variation in the dN/dS ratio and recombination rate (Green 1995; McVean *et al.* 2004). In this section I will focus on the differences in inference based on the new model and existing inference methods, including the advantages of the Bayesian approach. I will discuss the important assumptions of the model, its limitations for inferring natural selection and future extensions to the model. The implications of such methods for vaccine research are

discussed briefly, and finally I will discuss a separation of timescales approach for allowing intra-host genetic diversity without invoking coinfection as an explanation.

6.2.1 Comparison of PorB3 analyses

On the supposition that the analysis using the new PAC method is the most faithful account of the evolutionary history of selection and recombination in the *porB3* locus, then analyses based on comparison of the observed number of pairwise synonymous and non-synonymous differences (Smith *et al.* 1995) seriously underestimate the true extent of diversifying selection in the antigenic locus. Smith *et al.* (1995) estimated that the average dN/dS ratio was 0.62 for *porB3*, whereas in section 5.1.6 it was estimated to be 0.90 on average (under Prior A for the carriage study). Smith *et al.* (1995) estimated that the number of non-synonymous relative to synonymous substitutions in loop regions was 2.3, and 0.28 in non-loop regions. In the analysis presented in section 5.1.6 the average dN/dS ratio was estimated to be around 6 in loops I, V, VI and VII, and 0.16 elsewhere. So not only was the average dN/dS ratio underestimated, but the extent of the differences between positively and negatively selected sites was also greatly under-estimated. This is partly because some loop regions do not experience positive selection. As a result the analysis undervalues the importance of PorB as a potential vaccine target.

Using the phylogenetic CODEML method, Urwin *et al.* (2002) estimated an average dN/dS ratio of 0.26 for *porB3*, which is considerably smaller than the 0.90 estimated using the new method. Most of the difference probably lies in the way that the new method allows adjacent sites to share a common selection parameter, and that the new method also models insertions/deletions; sites segregating for an indel are deemed to

exhibit non-synonymous polymorphism, whereas in CODEML these sites are excluded from the analysis. For sites identified as experiencing strong positive selection, CODEML estimated a dN/dS ratio of 13.9, which is considerably higher than estimates of around 6 using the new method. This might in part reflect the constraining effect of the exponential prior on ω, which disfavours values of ω far from 1. However, it may also reflect the model misspecification that CODEML suffers from when the genes have undergone recombination. The tendency for phylogenetic methods to infer hypermutability at sites with homoplasies caused by recombination, discussed in section 5.3, may artificially elevate the inferred value of dN/dS.

CODEML and the new method differ substantially in the inferred patterns of variation in ω spatially along the gene. The distribution of positively selected sites inferred by CODEML is distinctly sporadic (see Chapter 5 Figure 7), whereas the new method generally infers much smoother variation in the mode of selection. This is a direct result of the prior on variation in ω, which models blocks of contiguous codons which share a common selection parameter, and the fact that the new method is able to infer selection at sites segregating for indels, which CODEML does not. The prior model of variation in ω is used deliberately to create a smoother posterior and share information between adjacent sites. The smoothness can be controlled by the prior value of the parameter p_ω. However, the length of a block follows a truncated geometric distributed, so very short blocks consisting of only a single codon have the highest probability mass under the prior. Therefore I would argue that if the data did not support smooth variation in the selection parameter along the sequence then the posterior would not be smooth. CODEML also infers positive selection at single sites

in loop II, loop IV, between loops IV and V, and two sites between loops V and VI, which the new method does not agree with. For the same reason that I do not believe the smoothness of the variation in ω within loops I, V, VI and VII is an artefact of the prior, I do not believe that the new method has failed to pick these sites out because of over-smoothing. Instead some at least may be false positives, caused by the assumption of no recombination in CODEML (see section 5.3).

6.2.2 Aspects of the Bayesian approach

Besides the ability to co-estimate ω and ρ, there are several advantages to the new method. Some of these are a consequence of the Bayesian approach, and all of them rely on the computational tractability of the PAC model. First among these is that the posterior probabilities of diversifying selection are fully Bayesian, so they incorporate uncertainty about the evolutionary history, as well as uncertainty in the other parameters. In the presence of recombination, there is likely to be a great deal of uncertainty in the evolutionary history. The computationally efficient PAC likelihood means that in the posterior, ω can take on any positive value, rather than having to constrain it to a discrete number of points or approximate a continuous distribution in a similar manner.

The main objection to a Bayesian approach is the requirement to specify a prior distribution on all parameters. In a scientific context it may seem absurd to prejudice the outcome of statistical inference with the researcher's prior subjective beliefs. In practice it is possible to represent a lack of prior knowledge with relatively flat priors, such as the proper and improper uniform priors used in ABC-CDE in Chapters 2 and 3, although it should be noted that in reversible-jump MCMC it is not possible to use

improper priors (Green 1995). In Chapter 5 I took a different approach, that of prior sensitivity analysis. Prior sensitivity analysis reveals which aspects of the posterior distribution, if any, are unduly influenced by the choice of prior. This in turn reveals which aspects of the model the data are uninformative about. For example, Chapter 5 Figure 6b shows that the data contained very little information about recombination rates at the extremes of the sequence. In contrast, inference about diversifying selection (Chapter 5 Figure 7) was robust to the prior.

In a Bayesian setting it is entirely natural to impose a block-like structure on the joint distribution of ω across sites. At sites where the data is compatible with a block-like structure this allows information about ω to be combined across sites, but when the signal in the data is strong enough it will overwhelm the block-like model. The sensitivity to the signal is controlled by p_ω. This is a biologically realistic model insofar as adjacent sites in the primary sequence will be closely juxtaposed in the tertiary structure, and, as such, are more likely to perform similar functional duties. If anything, the model is overly simplistic because the tertiary structure could in principle be used to impose longer-range dependencies on the prior. In a maximum likelihood setting, implementing the block-structure described here would be computationally unfeasible.

On the basis of previous work (Schierup and Hein 2000; Shriner *et al.* 2003; Anisimova *et al.* 2003) and because of clear model misspecification I have claimed that it is inappropriate to analyse data that shows evidence for recombination using phylogenetic methods. Yet neither the coalescent, nor the approximation to the coalescent used in Chapters 4 and 5, inevitably fit data from a recombining

population. That is why the importance of goodness-of-fit testing has been emphasised. Posterior predictive *p*-values allow for goodness-of-fit testing in a Bayesian setting when there is no explicit alternative model specified. The posterior predictive *p*-values in Chapter 5 Table 4 showed that the model with no recombination is a very poor fit to the data, and Chapter 5 Figure 7 showed that in the PAC model the assumption of no recombination leads to an increase in the number of sites experiencing diversifying selection, which would be expected if this assumption increases the false positive rate.

Posterior predictive *p*-values (Chapter 5 Table 4) suggested that the coalescent approximation was not a good fit to the *N. meningitidis* global study. This was not unexpected because the global study did not represent a random sample from any population in a meaningful sense. In constructing the carriage study care was taken not to include more than one haplotype from any one host. The idea was to envisage the bacterial population as a metapopulation, as described in Chapter 1. Consistent with this model, the posterior predictive *p*-values showed that the coalescent approximation did provide an adequate fit to the carriage study. There is more work to be done on formalizing the relationship between genetic models, such as the coalescent, and epidemiological models, but it may be possible in future to use models such as the one presented here to estimate parameters of epidemiological relevance. This is discussed further in section 6.2.6 below.

6.2.3 Limitations of the method

Fundamentally, the likelihood model used for inference is an approximation to the coalescent, and in that sense it is a compromise. In Chapter 1 I discussed why the

coalescent model is an appropriate null model for the evolution of microparasites such as *N. meningitidis*. The coalescent is a model of the genealogy of a random sample of genes in a selectively neutral population. However, integrating over the many possible, unknown, genealogies is computationally unfeasible for all but the simplest problems. The PAC model (Li and Stephens 2003) attempts to perform this integration implicitly, in a computationally convenient, but approximate, fashion. There are several trade-offs in this approximation. Broadly speaking, the biggest disadvantage is that there is no formal relationship between the coalescent and the PAC model, therefore it is very difficult to predict how the PAC model will behave differently to the coalescent. More specifically, one major problem is that the ordering in which the conditional likelihoods are calculated influences the likelihood, so that haplotypes are no longer exchangeable. For any reasonable number of sequences n, the number of possible orderings $n!$ is too large to fully explore, so the likelihood must be calculated using a finite number of orderings. Following Li and Stephens (2003) the new method averages over a fixed number of random orderings to calculate the likelihood. In contrast, Stephens and Scheet (2005) treat the ordering as a model selection problem, and integrate over the orderings numerically using MCMC. In hindsight, this might be a more elegant solution to the problem.

The PAC model is an approximation to a sampling formula (in the sense of Ewens [1972]) for a finite-sites mutation model in the presence of recombination. The $(k + 1)$th haplotype is a copy of the first k, but recombination means that it may be a mosaic, and mutation means it may be an imperfect copy. Recombination causes mosaicism because adjacent sites are more likely to share the same evolutionary history than a random pair of sites. In section 4.2 I argued that the haplotype from

which the (k + 1)th copies can be thought of the nearest neighbour in the genealogy at that site. In calculating the emission probabilities for the HMM, the time to the mrca of the (k + 1)th haplotype and its nearest neighbour is integrated out marginally for each site. The probability distribution is exponential with rate k to the order of the approximation. However, when adjacent sites share the same nearest neighbour, they are more than likely to share the same time to the mrca with that neighbour. Therefore integrating out the time marginally for each site is inconsistent with the coalescent model. In fact adjacent sites could share the same time to the mrca if time is discretized to retain the structure of the HMM (Stephens and Scheet 2005), but this increases the time to calculate the PAC likelihood. Discretizing time using Gaussian quadrature was the basis of the original importance sampler of Fearnhead and Donnelly (2001) that motivated the model of Li and Stephens (2003). Further investigation is needed to discern whether the additional complexity of discretizing time in the PAC likelihood would improve inference in the context of inferring selection and recombination rates.

Even if it were possible to use the actual coalescent model for the genealogical history of the gene sequences, the utility of the method presented here would remain limited by the biological realism of the mutation model. The NY98 mutation model is a useful way to treat selection that is confounded with mutation in samples of gene sequences. However, in Chapter 4 I argued that the ability of the model to detect positive directional selection in which one functionally constrained form is replaced by another is seriously questionable. The NY98 model may be best put to use in the context of inferring positive diversifying selection in genes that interact with the host immune system, because positive selection in the NY98 model is really selection for

genetic novelty. Detecting single adaptive events is best left to other methods that exploit other signals of selection in the data, such as strong haplotype structure and lowered diversity surrounding a site that has undergone a selective sweep (e.g. Przeworski *et al.* 2003; Coop and Griffiths 2004). PAC models may still be a useful approximation to the coalescent in this context.

6.2.4 Extensions to the method

There are several obvious extensions to the method, some of which would be relatively easy to implement. Among these is the replacement of the mutation model with any other reversible nucleotide, codon or amino acid mutation model. One interesting avenue that might be especially useful in the study of microparasites is to adapt the PAC likelihood to model genes sampled at different points in time. Serially-sampled genetic data is potentially a very powerful resource, because (i) genes sampled further back in time are informative about the genealogy at that time, (ii) serially-sampled data allows the effective population size to be deconfounded from the estimates of the mutation and recombination rates (Drummond *et al.* 2003a; Drummond *et al.* 2003b) and (iii) changes in selection pressures over time might be modelled. In this sense, microparasites are measurably evolving populations, because the mutation and recombination rates are sufficiently high and generation length sufficiently short for their evolution to be observed in real-time (Drummond *et al.* 2003b). This might be interesting not only from an evolutionary perspective, but could be instructive for control and prevention of emerging infections or epidemics.

In contrast to the problem of incorporating serially-sampled data into a PAC likelihood, a straightforward extension to the method would be to allow more than

one isolate per host to be included in the sample. Provided that the origin of each of the isolates is known, Wakeley and Aliacar (2001) provide a genealogical model for repeated sampling of some hosts that could be easily incorporated into the model presented here. In essence, the PAC model used for inference in this chapter models the collecting phase of Wakeley and Aliacar's metapopulation genealogy. This extension would allow their scattering phase to be modelled as well. Incorporating the scattering phase into the model would involve a variable number of haplotypes at the beginning of the collecting phase, which could be integrated out as part of the MCMC scheme. In Wakeley and Aliacar's model (see section 1.4), sequences sampled from the same host either coalesce with others sampled from the same host or migrate (backwards-in-time, as a result of a transmission event) to a new host, before commencing the collecting phase. From the perspective of inference, only sequences that are identical can coalesce during the scattering phase if the mutation rate is finite on the timescale of the collecting phase, which implies that sequence variation within a host can only be explained by multiple infection under the model. In section 6.2.6 I discuss why this might not need to be the case.

6.2.5 Implications for vaccine research

Identifying sites in an antigen locus that are under positive selection can help to locate the determinants of antigenicity, because interaction with the host immune system causes selection for antigenic novelty, which is brought about by variation in the amino acid sequence. It has been claimed that identifying the determinants of antigenicity might inform vaccine research (de Oliveira *et al.* 2004). Currently meningococcal vaccines for non-serogroup-B meningococcal disease use the capsular polysaccharide or a conjugate polysaccharide-protein complex (Stuart 2001). Genetic

analysis such as that presented here cannot inform such studies because there is not a one-to-one correspondence between the nucleotide sequence and the antigen (the protein in the conjugate vaccine is only a carrier). However, much current research in vaccine development is focussing on serogroup B (Snape and Pollard 2005), for which no efficacious vaccine currently exists, owing to the poor immunogenicity of the serogroup B polysaccharide (see section 1.1.3). An alternative to conjugate polysaccharide vaccines are outer membrane vesicle (OMV) vaccines which can be readily obtained from the blebs constantly secreted by the outer membrane. The outer membrane proteins (OMPs) that are the immunodominant components of the OMV vaccines can also be synthesised in the laboratory, but it is difficult to simulate the conditions required for the natural conformation of the proteins (Frasch 1995).

In the context of OMV vaccines, genetic analyses are potentially of use. On the premise that positive selection in known antigens reflects interaction with the immune system, then the strength of interaction can be quantified using the selection parameter, either at a particular site or summed across sites. This could direct research into the protein components of the OMV and find the genetic determinants of immunodominance. When combined with knowledge of the tertiary structure of the proteins in their natural conformation, this might prove to be a useful tool. Of particular interest might be, not sites that are demonstrably under strong diversifying selection, which in this study were found to be those that were surface-exposed, but rather those that maintain strong functional constraint despite a prominent surface-exposed position. Such a result might suggest that the conservation of the residue or oligopeptide is so essential to the correct functioning of the protein that it is constrained despite strong selection for antigenic variation. Generally speaking, the

technology of OMV vaccines is currently too crude for fine mapping of immune selection on a gene sequence to be of great use. However, genetic analyses such as that presented here are inexpensive in comparison to laboratory experiments or field trials. This alone is a persuasive argument for pursuing population genetic analysis in the context of vaccine research even if the rewards are slight.

6.2.6 Separation of timescales in microparasite evolution

In Chapter 1 a metapopulation was used to model a population of hosts infected with a microparasite, where the epidemiological dynamics are described by a simple SIRS-style differential equation model. Using the coalescent model of a metapopulation (Wakeley and Aliacar 2001) provides a starting point for modelling the genealogy of gene sequences sampled from a microparasite population. When all isolates are sampled from different hosts, the genealogy is simply the coalescent, where the effective population size is a function of the epidemiological parameters. When some isolates are sampled from the same host, the genealogical history has two phases, the scattering phase and the collecting phase (Wakeley and Aliacar 2001). In the scattering phase the lineages ancestral to the sample coalesce within each host, or migrate backwards-in-time to other hosts. This occurs rapidly relative to the subsequent collecting phase, which is a coalescent process with altered time scale. In Wakeley and Aliacar's model, if the mutation rate in the collecting phase is finite, then no mutation events occur in the scattering phase because it occurs so rapidly relative to the collecting phase. The same is true of recombination. This implies that genetic variation in the parasite population within a single host must be caused by coinfection.

In reality there may be appreciable genetic variation in a host even if the infection had a single founder. This poses a problem for the genealogical model because mutation would then occur infinitely quickly during the collecting phase. It is trivial to show that this is untrue simply by examining a sample in which each host is represented by only a single isolate. One explanation is that the genealogical model is wrong. The key assumption is that there are a large number of demes, which allows the separation of timescales into a scattering and collecting phase. This assumption, however, seems reasonable. The separation of timescales is a very convenient result, and it might be premature to abandon it at this stage. In addition to the assumption that the within-host population size is large $(N_P \to \infty)$ and the number of hosts is large $(D \to \infty)$, Wakeley and Aliacar assume that $D >> N_P$. For viral microparasites such as HIV, clearly this assumption may not hold. However, it may be possible to observe intra-host variation without invoking multiple infection even when $D >> N_P$ is a reasonable assumption, using the same idea as the NY98 mutation model that mutation and selection are confounded.

During transmission, there is a bottleneck in genetic diversity. The bottleneck in diversity is caused by selection for genotypes that are adapted to transmission. Put another way, upon colonisation of a host, there is a relaxing of the selection pressure allowing the population to diversify. In any parasite population, only a fraction of genotypes will be competent for transmission. In a model such as NY98, selection is confounded with the mutation process. This is useful in the context of resolving the conflict between intra-host and inter-host genetic diversity. To illustrate the point, assume that the mutation rate μ within a gene is finite on the timescale of the scattering phase (explaining intra-host diversity) and that within the host there is no

selection acting on the gene; it is neutral. If that gene is important for transmission, so that only a fraction f of all forms are competent for transmission, then the effective mutation rate in an inter-host sample will be $f\mu$. If f is sufficiently small so that $O(1/f) > O(\mu)$, then the effective mutation rate might be finite on the timescale of the collecting phase, so that the gene would exhibit intra-host diversity yet also inter-host genetic structuring.

The idea of an effective mutation rate, where selection is modelled as a form of mutational bias, is the essence of the NY98 model. This suggests that intra-host and inter-host selection pressures should be separately parameterised, which is biologically reasonable because the adaptations required for surviving in an infected host may be quite different to those required for successful transmission to a new host, and there may be limited overlap between the two. Inference of intra-host and inter-host selection pressures could be performed in the context of the model presented here, when extended to incorporate the scattering phase of the genealogical process. In the terminology of Wakeley and Aliacar (2001), it would require integration (by MCMC) over the sample configuration at the end of the scattering phase. In principle, intra-host and inter-host selection maps could be co-estimated using the method presented here as the foundation.

6.3 Summary

Patterns of genetic diversity in parasite populations contain an, albeit corrupted, account of the evolutionary history of the population. Understanding that evolutionary history can help inform control and prevention strategies for pathologically

importance parasites such as *N. meningitidis*. The large genetic diversity, short generation times and intimate co-evolutionary interaction between host and parasite also make pathogens interesting case studies in the study of evolution. The right way to model genetic data is to take account of the strong inter-dependency of gene sequences imposed by the evolutionary tree. For gene sequences sampled at random from a population, the coalescent provides the appropriate null model (or prior distribution) for the underlying evolutionary tree, or trees in the presence of recombination. Evolutionary models for genetic data are the only way to obtain biologically relevant and interpretable parameter estimates. Improving understanding of the biology of pathogens by iteratively refining and criticizing the model of the evolutionary ancestry, arguably the most important role for mathematical models in evolution, can only be achieved by using biologically interpretable explicit models of evolution. Even so, descriptive methods such as AMOVA and Mantel tests have a role for exploring genetic datasets and generating hypotheses. Because of the inherent computational difficulties in performing inference on evolutionary models, further research is required into approximate techniques such as the PAC model and techniques based on summaries of the full data such as ABC-CDE. Used in combination, there is the potential to learn a great deal about the evolution of the microorganisms responsible for important infectious diseases of humans.