J Mol Evol (2005) 61:90–98
DOI: 10.1007/s00239-004-0290-6

JOURNAL OF MOLECULAR EVOLUTION

Evidence of Positive Darwinian Selection in Putative Meningococcal Vaccine Antigens

David A. Fitzpatrick, Christopher J. Creevey, James O. McInerney

Department of Biology, National University of Ireland, Maynooth, Co. Kildare, Ireland

Received: 24 September 2004 / Accepted: 25 February 2005 [*ReviewingEditor:* Rasmus Nielsen]

Abstract. Meningococcal meningitidis is a life-threatening disease. In Europe and the United States the majority of cases are caused by virulent meningococcal strains belonging to serogroup B. Presently there is no effective vaccine against serogroup B strains, as traditional vaccine antigens such as polysaccharide capsules are unusable as they lead to autoimmunity. The year 2000 saw the publication of the complete genome of *Neisseria meningitidis* MC58, a virulent serogroup B bacterium. Working in conjunction with the sequencing project, researchers endeavored to locate highly conserved membrane-associated proteins that elicit an immune response. It is hoped that these proteins will provide a basis for novel vaccines against serogroup B strains. A number of potential vaccine antigens have been located and are presently in phase I clinical trials. Recently many reports pertaining to the evidence of positive Darwinian selection in membrane proteins of pathogens have been reported. This study utilized *in silico* methods to test for evidence of historical positive Darwinian selection in seven such vaccine candidates. We found that two of these proteins show signatures of adaptive evolution, while the remaining proteins show evidence of strong purifying selection. This has significant implications for the design of a vaccine against serogroup B strains, as it has been shown that vaccines that target epitopes that are under strong purifying selection are better than those that target variable epitopes.

Correspondence to: James O. McInerney; *email:* james.o.mcinerney@may.ie

Key words: *Neisseria meningitidis* — Positive selection — Vaccine design

Introduction

Neisseria meningitidis is a gram-negative, encapsulated β-proteobacterium that is naturally competent for transformation with DNA. It only thrives in the human host and is not known to colonize any other animal or environmental niches. Meningococcal carriage is very much more common than disease. Asymptomatic colonization of the nasopharynx is common, reaching 10% or more of the population in many countries. However for reasons still poorly understood, certain strains can penetrate the mucosal epithelium and gain access to the circulation system (Grandi 2003). In individuals lacking humoral immunity to meningococci, proliferation of the organisms in the blood may lead to septicaemia, characterized by circulatory collapse, multiple organ failure, and coagulopathy, additionally some virulent strains may also reach the subarachnoid compartment and initiate meningitis (Nassif et al. 1999). Meningococcal meningitis is a significant public health problem and is responsible for deaths and disability through epidemics in sub-Saharan Africa, and many other sporadic cases worldwide (Tettelin et al. 2000). Without prompt antibiotic treatment, meningococcal infection is almost always fatal (Nassif 2002), and even with prompt treatment death and sequelae are common (Naess et al. 1994). Those

most at risk from meningococcal infection include persons in the 15–24 age group and children under the age of 5. Annual incidence of meningococcal disease varies from 0.5 to 10 per 100,000 persons, but during epidemics the incidence rate can rise to above 400 per 100,000 persons. Presently five pathogenic *N. meningitidis* serogroups (A, B, C, Y, and W135) have been described based on capsular polysaccharide typing (Gotschlich et al. 1969).

To date there has been some success in vaccine design against four of the five pathogenic serogroups (A, C, Y, and W135). These vaccines consist of purified polysaccharide antigens, are highly effective in adults and the high molecular polysaccharides used in these vaccines are produced in the same manner as first described by Gotschlich et al. (1969). There is presently no effective vaccine against serogroup B, however, as the capsular polysaccharide of this bacterium is identical to a widely distributed human carbohydrate (polysialic acid), preventing the use of this as a vaccine antigen for fear it may induce autoimmunity (Nassif 2002). Vaccine design against strains of serogroup B is a priority as they cause the majority of invasive disease in Europe ($>50\%$ of cases) and the United States ($>30\%$ of cases).

The problems associated with developing a meningococcal serogroup B vaccine has led researchers to explore the possibility of targeting membrane-exposed proteins. Administration of an immunizing agent early in life should induce the production of serum bactericidal antibodies, which are the mark of resistance of meningococcal disease (Goldschneider et al. 1969). Suitable antigens include those that are conserved among a population, are expressed on the surface of the bacteria, and also induce the production of bactericidal antibodies (Nassif 2002). Using these criteria, the complete genome sequence of *N. meningitidis B* (strain MC58) (Tettelin et al. 2000) was used to locate open reading frames (ORFs) that potentially encode surface-exposed or exported proteins (Pizza et al. 2000). From the 570 candidate ORFs located they successfully cloned and expressed 350 of these into *Escherichia coli*. The resultant recombinant proteins were used to immunize mice and the immune sera were tested for bactericidal activity, as this correlates with protection in humans (Pizza et al. 2000). Finally, from 85 proteins that elicit bactericidal activity the suitability of these proteins as candidate antigens for conferring protection against heterologous meningococcal strains was tested. The tests were carried out on 34 different *N. meningitidis* clinical strains isolated worldwide and over many years. This process resulted in the identification of seven proteins that confer protection against both homologous and heterologous meningococcal strains. The sera against these antigens are capable of killing all the meningococcal strains so far utilized in the complement-mediated bactericidal assay, thus making the antigens particularly promising for vaccine formulations. Presently phase 1 clinical studies are in progress to establish the ability of these antigens to induce bactericidal antibodies in humans (Grandi 2003). These genes are highly conserved, which is an unusual finding considering that in three decades of studies, with one exception (Martin et al. 1997), only antigenically variable surface exposed proteins had been described (Pizza et al. 2000). Furthermore, the frequency of recombination of these genes is relatively low (0.011 using the homoplasy test [Maynard-Smith and Smith 1998]), a level of recombination that is similar to that of *Neisseria* housekeeping genes (Maynard-Smith and Smith 1998).

Recently a number of membrane-associated meningococcal proteins (Andrews and Gojobori 2004; Smith et al. 1995) have been shown to be under the influence of positive Darwinian selection (adaptive evolution). Adaptive evolution is a process that encourages the retention of mutations that are beneficial to an individual or population (Creevey and Mclnerney 2002). In protein coding genes, positive selection is thought to be an ephemeral event frequently leading to the generation of novel protein function (Kinsella et al. 2003). At the DNA level, positive selection may be detected by comparing the rate of nonsynonymous (amino acid altering) nucleotide substitutions per nonsynonymous site (d_N) with that of synonymous substitution per synonymous site (d_S). When $d_N/d_S > 1$ positive selection is said to be operating on the genes in question, alternatively when $d_N/d_S < 1$ purifying selection is said to be operating. Positive selection has been observed in membrane-associated genes of several pathogens (Andrews and Gojobori 2004; Bush et al. 1999; Fares et al. 2001; Jiggins et al. 2002; Kinsella et al. 2003; Urwin et al. 2002; Yang 2000; Yang 2001; Zanotto et al. 1999) and can be described as a host–parasite arms race (Jiggins et al. 2002). As the membrane proteins of pathogens are exposed to the host immune system, it follows that they may be under the greatest selective pressure for change. From mathematical modeling of viral infection dynamics it has been suggested that conserved epitopes are more appropriate as vaccine targets than variable epitopes (Nowak et al. 1991). Therefore, it would be helpful if we could predict candidate epitopes computationally, as it should accelerate the entire vaccine process. Observations have linked the success and failure of present-day vaccines to the presence/absence of positive selection. For example, annual vaccines against influenza A virus has to be developed due to its rapid antigenic change; conversely, the extremely successful poliovirus vaccine has been shown to be under strong negative selection (Suzuki 2004). These observations

would lead us to conclude that candidate vaccine targets should contain no positively selected sites, a view that has been expressed by others (Suzuki 2004).

From a protein structure perspective, it is likely that there are two categories of proteins, those that are amenable to change and those that, for structural reasons, cannot easily change. An ideal vaccine candidate should be in the second category and it should be possible to determine from analyzing the evolutionary history of a set of sequences if change via positive selection is easy or difficult in those sequences. In this paper we test the hypothesis that the seven gene families currently being considered as potential vaccine candidates (Pizza et al. 2000) show evidence of historical positive selection. The reason for proposing this situation is that these proteins are known to be expressed at high levels, are surface exposed, and elicit a strong immune reaction during infection. We discuss our findings in light of the implications for vaccine design.

Materials and Methods

Sequence Data

Twenty-two *N. meningitidis* serogroup B strains, three serogroup A strains, two serogroup C strains, one strain each of serogroups W, X, Z, and Y, one strain of *N. cinerea*, one strain of *N. lactamica*, and three strains of *N. gonnorrhoeae* were used in this analysis. Seven gene families (putative vaccine targets) were examined. The nomenclature of each family corresponds to the annotation of the completely sequenced MC58 genome (Tettelin et al. 2000), Therefore, we call these families by the names NMB0033, NMB0992, NMB1162, NMB1220, NMB1946, NMB2001, and NMB2132. The accession numbers for each of the genes is as follows: NMB0033 (AF226387–AF226417 and AF235143–AF235145), NMB0992 (AF226356–AF226386), NMB1162 (AF226542–AF226565 and AF235157–AF235158), NMB1220 (AF226542–AF226572 and AF235157–AF235159), NMB1946 (AF226480–AF226510 and AF235151–AF235153), NMB2001 (AF226449–AF226479 and AF235148–AF235150), and NMB2132 (AF226418–AF226448 and AF235146–AF235147). See Pizza et al. (2000) for a complete list of references.

Terminal stop codons were removed from all sequences. Any nucleotide sequence that was found to be identical to another strain was removed from the analysis. The nucleotide sequences were translated into their amino acid equivalents and aligned using CLUSTALW verl.82 (Thompson et al. 1994). Gaps created in the amino acid alignment were transposed back to the nucleotide sequences to gain a codon-based alignment using the program Putgaps (http://bioinf.may.ie/software/putgaps). Codon alignments were corrected for obvious alignment ambiguity using the alignment editor Se-Al 2.0a11 (http://evolve.zoo.ox.ac.uk/software.html?id=seal). The average pairwise similarity for all gene families was found to be greater than 95%.

Data Analysis

Phylogenies for all gene families were constructed using the maximum likelihood criterion in PAUP 4 (Swofford 1998), the optimal model of sequence substitution was selected for by comparing the likelihood scores using MODELTEST 3.04 (Posada and Crandall 1998).

Using the method of Gassly and Holmes (1997) as implemented in PLATO 2.0, we searched for blocks of sequences that have incongruent phylogentic topologies due to recombination. PLATO locates spatial variation in an alignment that results in different regions of the alignment supporting different phylogenies. Such variation is indicative of recombination. PLATO utilizes the maximum likelihood phylogenetic tree for a gene together with its substitution model and calculates the likelihood of this hypothesis along the alignment. The reasoning behind this step is that recombination leads to apparent substitution rate heterogeneity and can closely resemble the effects of molecular adaptation, thus by excluding incongruent blocks we will lessen the possibility of finding false positive results.

In the absence of tertiary structure data, the ConSeq (Berezin et al. 2004) web server was used for the identification of biologically important residues within all seven gene families. Functionally important residues are often solvent accessible and evolutionary conserved while structurally important residues are normally highly conserved and found within the protein core. Ideally, vaccine targets should be from the former category, as they are accessible and not likely to change. Using ConSeq we hope to locate such exposed functionally important residues.

Analysis of Selection

The likelihood ratio test (LRT) approach of Yang et al. (2000) as implemented in the PAML package 3.13 (Yang 1997) was used to examine selection pressures acting on different amino acid positions in the meningococcal gene families. Given that selective pressures are likely to vary across different sites in a protein sequence, models have been developed to incorporate heterogeneous selective pressures at different sites (Yang et al. 2000). This approach evaluates nested models of sequence evolution. Some models are more parameter-rich extensions of other models, and when this is the case, an LRT may be performed with twice the log-likelihood difference being compared with a χ^2 distribution with the degrees of freedom equal to the difference in the number of parameters. We used three LRTs. The first compares a model that assumes one $d_N{:}d_S$ ratio (ω hereafter called model M0) for all sites with a model that assumes two site classes with independent Δ values estimated from the data (hereafter called model M3; $k = 2$). The second LRT compares a model that allows two site classes with values fixed at 0 and 1 (hereafter called M1 [neutral]), with a model that has an additional site class that allows ω to be greater than unity (hereafter called model M2 [selection]). The final LRT compares M7, which has ω β-distributed between 0 and 1 among 10 site classes each of equal proportion, with a model that allows an additional site class where ω is freely estimated (hereafter called model M8). The comparison of M7 *vs* M8 is the most powerful test of positive selection (Anisimova et al. 2001).

The second major step identifies those codon sites under positive selection when the LRT suggests their presence. This is achieved using the Bayes theorem to calculate the posterior probabilities sites are from different ω classes (Nielsen and Yang 1998). Positions with a high probability of coming from the class with $\omega > 1$ are likely to be under positive selection (Swanson et al. 2001).

Maximum likelihood models that allow for heterogeneity in the d_N/d_S ratio among lineages were also tested. The simplest model is the one-ratio model, M0. The most general model is the free-ratio model, which assumes as many ω parameters as the number of branches in the tree; this is a parameter-rich model (Yang 1998). Models that fit between these two extremes include the two-ratio model and three-ratio models; these allow predefined lineages to have a different ω value from the rest of the tree (Yang 1998). All of

Table 1. ML analysis of NMB0992 using a variety of models

Model	Estimates of parameters	InL	2Δln$L$	Positively selected codons
M0 (one ratio)	$\omega_1 = 0.3868$	−3396.06		None
M1 (neutral)	$p_1 = 0.81$, $\omega_1 = 0.00$ $p_2 = 0.18$, $\omega_2 = 1.00$	−3326.86		Not allowed
M2 (selection)	$p_1 = 0.87$, $\omega_1 = 0.00$ $p_2 = 0.02$, $\omega_2 = 1.00$ $p_3 = 0.09$, $\omega_3 = 3.95$	−3298.77	(M1 *vs* M2) 56 $p < 0.0005$	18, **35**, 44, 50, 62, 64, 65 67, 78, 83, 102, 149, 176 278 373, 417, **472**
M3 (discrete $k = 2$)	$p_1 = 0.88$, $\omega_1 = 0.00$ $p_2 = 0.11$, $\omega_2 = 3.67$	−3298.84	(M0 *vs* M3) 194 $p < 0.0005$	11, 18, **35**, 44, 50, 51,59, 61, 62, 64, 65, 67, 70, 74, 78, 83, 94, 102, 120,135, 137, 139, 143, 149, 176, 200, 202, 240, 254, 261, 278, 338, 351, 372, 373, 374, **382**, **383**, 417, 426, 439, 454, **472, 476**, 491
M7 (β)	$p = 0.00130$ $q = 0.00455$	−3327.02		Not allowed
M8 (β & ω)	$p = 0.001$, $q = 1.81$ $p_2 = 0.11$, $\omega_1 = 3.67$	−3298.84	(M7 *vs* M8) 56 $p < 0.0005$	11, 18, **35**, 44, 50, 51,59, 61, 62, 64, 65, 67,70,74, 78, 83, 94, 102, 120,135,137, 139, 143, 149, 176, 200, 202, 240, 254, 261, 278, 338, 351, 372, 373, 374, **382**, **383**, 417, 426, 439, 454, **472**, **476**, 491

Note. p_1, p_2, and p_3 refer to the proportion of sites in categories 1, 2, and 3, respectively. ω refers to the d_N/d_S ratios in these categories of sites. p and q are β estimates. Models M1 and M7 do not allow positively selected sites. The degrees of freedom for all likelihood ratio tests is 2. Positively selected codons in boldface represent the codons that are found by both the ML and the MP method as being under the influence of positive Darwinian selection. For the ML analysis, only sites with a Bayesian probability greater than 95% are considered significant, while a 95% confidence level was also assigned to the MP analysis.

these methods were utilized in an effort to detect if positive selection has acted along major lineages within the seven gene families.

Recent studies have shown that under certain conditions ML methods can be sensitive to violations of assumptions made in models that test for positive selection; these sensitivities under certain conditions can result in false positives (Suzuki and Nei 2001a, 2004). To account for this, a maximum-parsimony (mp) method that tests for adaptive evolution was applied. A sliding window procedure that uses the model of Li (1993) as implemented in SWAPSC (Fares 2004) was used. This method infers a statistically optimum codon-window size and slides it along the alignment. Each window step is then tested for the significance of nonsynonymous substitutions to synonymous substitutions and the nonsynonymous-to-synonymous rate ratio ω; in this manner, positively, neutrally, or negatively evolving sites can be located. This method also has the ability to test for saturation of synonymous substitutions (Fares 2004); if any sites are highlighted, they can be removed.

Results

Analysis of Recombination

The effects of recombination on methods that detect positive selection have been documented (Anisimova et al. 2003). In an effort to ensure that these pitfalls do not affect the results reported, the method of Grassly and Holmes (1997) was utilized to ensure that no recombination had occurred within our dataset. This method analyzes an alignment for sequence blocks within the alignment that deviate significantly from a predefined topology (the ML tree found earlier). No blocks within the vaccine candidate genes were found to deviate significantly from the imposed phylogenies. Saturation of synonymous sites is also an important issue when trying to detect adaptive evolution events by the criterion that d_N/d_S is greater than 1, as saturation can lead to the underestimation of d_S and an inflation of the d_N/d_S ratio. The moving window program SWAPSC (Fares 2004) was used in an effort to test for saturation of synonymous sites. The vaccine candidate genes were shown not to contain saturated synonymous sites, and for all seven gene families more than 95% of sites show strong evidence of strong purifying selection.

Maximum Likelihood Analysis

Five of the seven vaccine candidate genes (NMB0033, NMB1162, NMB1220, NMB1946, and NMB2001) did not exhibit strong evidence of positive selection (results not shown). Gene families were only placed into the category of undergoing positive selection if all three LRTs performed were significant. Analysis of the selective pressures acting on the remaining two families (NMB0992 and NMB2132) provided strong evidence for positive selection (Tables 1 and 2). The three LRTs for the NMB0992 data indicated that there was strong evidence for positive selection at some sites. All three LRTs performed were highly significant (Table 1). M2 indicated that approximately 9% of sites had an ω value of ~3.95. For both M3 and M8 approximately 11% of sites fell into a strong positively selected class ($\omega = 3.67$). The positively selected codons identified by Bayesian posterior probability were identical for M3 and M8. A

Table 2. Results of the ML analysis of NMB213z using a variety of models

Model	Estimates of parameters	ln*L*	2Δln*L*	Positively selected codons
M0 (one ratio)	$\omega = 0.57$	−6827.19		None
M1 (neutral)	$p_1 = 0.58$, $\omega_1 = 0.00$ $p_2 = 0.42$, $\omega_2 = 1.00$	−6609.20		Not allowed
M2 (selection)	$p_1 = 0.56$, $\omega_1 = 0.00$ $p_2 = 0.33$, $\omega_2 = 1.00$ $p_3 = 0.093$ $\omega_3 = 3.97$	−6552.92	(M1 *vs* M2) 435 $p < 0.0005$	**217**, **222**, **224**, **260**, **264**, **268**, **276**, 284, **290**, **293**, **305, 309, 311**
M3 (discrete $k = 2$)	$p_1 = 0.70$, $\omega_1 = 0.07$ $p_2 = 0.30$, $\omega_2 = 2.13$	−6570.05	(M0 *vs* M3) 514 $p < 0.0005$	**193, 217, 222, 224, 260, 264, 268, 276,** 284, **289, 290, 292, 293, 305, 309, 311**
M7(β)	$p = 0.018$, $q = 0.035$	−6612.12		Not allowed
M8 (β & Δ)	$p = 0.02$, $q = 0.03$ $p_1 = 0.11$, $\omega_1 = 3.66$	−6552.80	(M7 *vs* M8) 118 $p < 0.0005$	**193, 217, 222, 224, 260, 264, 268, 276,** 284, **289, 290, 292, 293, 305, 309, 311**

Note. p_1, p_2, and p_3 refer to the proportion of sites in categories 1, 2, and 3, respectively. Δ refers to the d_N/d_S ratios in these categories of sites. p and q are β estimates. Models M1 and M7 do not allow positively selected sites. The degrees of freedom for all likelihood ratio tests is 2. Positively selected codons that are boldface and underlined represent the codons that are found by both the ML and the AMP method to be under the influence of positive Darwinian selection. For the ML analysis only sites with a Bayesian probability greater than 95% are considered significant, while a 95% confidence level was also assigned to the MP analysis.

smaller number was suggested by M2, but these were subsequently found to be a subset of those predicted by M3 and M8. The LRT analysis of site-specific variable selective pressures acting on NMB2132 again provided support for hypotheses of adaptive evolution (Table 2). M0 estimated an average ω value of 0.579. Again, using models that allow for variable ω values across sites, all three LRTs are highly significant ($p < 0.0005$) and all three models find more than 9% (9.6, 30, and 10% for models M2, M3, and M8, respectively) of sites to have an ω value greater than 1 (3.97, 2.13, and 3.66). Bayesian posterior probabilities indicate that the same 16 codons can be assigned with confidence ($p > 0.95$) to this class of sites with a high ω value for M3 and M8 while a smaller number (13) are found for M2. These 13 sites overlap with those found by the other two models.

In order to test the hypothesis that one or more lineages are responsible for the finding of positive selection or to identify lineages where positive selection is stronger, likelihood models that allow for different d_N/d_S ratios among evolutionary lineages (Yang 1998) were used. In all cases none of the LRTs performed allowed for positive selection on any one lineage exclusively (results not shown). This indicates that the results from the first set of LRTs are the result of constant selection on specific sites across all lineages and not focused on any particular lineage.

Parsimony Analysis

A number of sites inferred to be under the influence of positive Darwinian selection using the ML method were not inferred using the parsimony sliding window approach (Tables 1 and 2). The parsimony method used in this analysis does not consider sites where d_S is equal to 0 as having undergone positive selection even if the d_N ratio is greater than 0. The ML methods identify situations such as this as positively selected events; therefore, the parsimony method is a more conservative approach. This is the reason why complete agreement between both methods is not observed. In this study, we treat these sites as being under the influence of positive selection. For NMB0992 the sites inferred to be under positive Darwinian selection have an average ω value of 3.95. An average ω value of 3.12 was found for NMB2132. The ML and MP methods that can detect positive selection are generally in good agreement with one another. Both methods inferred nearly the same set of codons as having undergone positive Darwinian selection for NMB213, and also for NMB0992 but to a lesser extent (Tables 1 and 2).

Functionally Important Residues

A large number of functionally important residues were predicted for both NMB0992 and NMB2132 by ConSeq (Figs. 1 and 2). Obviously, the sites that have been shown to be evolving under positive Darwinian selection do not fit into this category, as there must be a degree of variability at such a site before it can be inferred to have undergone positive selection. The positively evolving sites inferred by the ML method are nearly all exposed (i.e., are not coiled into the centre of the protein) and therefore most probably interact with the host immune system. It is not particularly surprising that all codons identified as being subject to positive selection encode amino acid acids exposed on the protein surface, as not only are these potentially subject to immune surveillance but also they are likely to be structurally less constrained.

```
MNKIYRIIWN SALNAWVVVS ELTRNHTKRA SATVETAVLA TLLFATVQAN
eeebbebbbb bbbebbbbbb ebbeebeeeb ebebebbbbb bbbebbbebe
fffssfssss  ssfsss ss fssffsffss fsfs sssss sss sssfs  50

AFTYSLKKDL TDLTSVGTEE LSFGANGNKV NITSDTKGLN FAKKTAGTNG
bbeeebeeeb eebeeeeeee bebeeeeeeb ebebeeebbe bbeeeeeeee
 sfffsff s   s  f ff  sfs fff fs fs sfffssf ssf ffffff 100

DTTVHLNGIG STLTDRAASI KDVLNAGWNI KGVKTGSTTG QSENVDFVRT
eeebebbbbb bebeeebeeb eebbebbbeb eeeeeeeeee eeeebebbbb
f fsfsssss sfsfffsff  ffssfsssfs ffff f f f ff fsfss s 150

YDTVEFLSAD TKTTTVNVES KDNGKRTEVK IGAKTSVIKE KDGKLVTGKG
bebbebbeee eeebebebeb eeeeeebebe bebeebbbee eeeeebeeee
sfssfssfff fffsfsfsfs fffff sfsf sfsffsssff fffffsfff  200

KGENGSSTDE GEGLVTAKEV IDAVNKAGWR MKTTTANGQT GQADKFETVT
eeeeeeeeee eebebbbeeb bebbbebebe beeeeeeeee eebeebeebe
f ffffffff ffsfsssffs sfsssfsfsf sffffffff  ffsffsffsf 250

SGTNVTFASG KGTTATVSKD DQGNITVKYD VNVGDALNVN QLQNSGWNLD
eeeebebbee eeeebeeeee eeeebbbeee bebbebbebe ebeeeeeebe
fff sfssff  fffsffffff ffffsss ff sfssfssfsf fsfffffffsf 300

SKAVAGSSGK VISGNVSPSK GKMDETVNIN AGNNIEITRN GKNIDIATSM
bebbeeeebe bbbbeeeeee eebeeebebe beeebebeee beebebbbbb
sfssfffffsf ssssffffff ffsfffsfsf sfffsfs ff sffsfsssss 350

TPQFSSVSLG AGADAPTLSV DDEGALNVGS KDANKPVRIT NVAPGVKEGD
beebbebbbe bbbebeebeb eeeebbebee eeeeeebebb ebbeebeeee
 ffssfsssf sssfsffsfs f   ssfsff f  fffsfss fssffsffff 400

VTNVAQLKGV AQNLNNHIDN VDGNARAGIA QAIATAGLVQ AYLPGKSMMA
beebbebeeb beebeeebee bebebeeebb ebbbebebbb bebeeebbbb
sfssfffffsf ssssff fff ffsff sfsf sfffsfsf f sffsfsssss 450

EAGYAIGYSS ISDGGNWIIK GTASGNSRGH FGASASVGYQ W
eebebbbeeb bebbbebebe beeeeeeeee eebeebeebe e
fff sfssss fssfsssfss fssff fffs fsfsssssss
```

Fig. 1. ConSeq predictions of structure for NMB0992 demonstrated on AF226356, using all homologous genes within this family. The sequence of the query protein is displayed on the first row. The second row lists the predicted burial status of the site (i.e., b, buried, versus e, exposed). The third row indicates residues predicted to be structurally and functionally important: s and f, respectively. Sites inferred as evolving under the influence of positive selection are in boldface and underlined; the majority of such sites are in exposed regions of the protein.

Discussion

The development of a vaccine against heterologous *N. meningitidis* B strains is an ongoing process. Despite more than 20 years of research (Bjune et al. 1991; Frasch 1989; Sierra et al. 1991), vaccines to protect against heterologous strains have yet to be developed. Candidate vaccine designs including the purification of the polysaccharide capsule are not an option, as this may lead to autoimmunity. Researchers are instead trying to locate highly conserved membrane proteins that elicit a host immune response. This analysis has found that two such vaccine candidates display evidence of positive selection. This finding may have implications for vaccine design, as a protein that has exhibited an ability to undergo selection in the past may do so again. While we can only speculate, it is plausible that further change at important antigen binding sites may make any vaccines developed from these proteins obsolete.

In this analysis, LRTs were used to model variable selective pressures across sites and lineages for a variety of vaccine candidates that are currently being tested. Recently concerns have been raised regarding the frequent false-positive results inferred using this ML approach (Suzuki and Nei 2001b, 2004). In an effort to add further validity to the ML results, a parsimony analysis was also performed. According to the ML analysis for selection within NMB2132, 16 sites are under the influence of adaptive evolution; the parsimony method infers 15 of these sites also. There were some discrepancies between the two methods when NMB0992 was analyzed, as the ML method inferred 45 sites, compared to the parsimony method, which only inferred 5 of these sites. Therefore we conclude that there is excellent evidence that 15 sites within NMB2132 have undergone positive selection, while there is corroborated evidence to suggest that at least five sites within NMB0992 have also undergone an adaptive event. Furthermore, the majority of these sites are found within exposed regions and therefore are likely to be in direct contact with host immune responses.

High levels of recombination can affect ML analysis as can saturation of synonymous sites (Anisimova et al. 2003). These concerns were carefully considered, yet in all seven gene families examined no signatures of major recombination events or saturation of synonymous sites were observed. Previously these seven proteins have been shown to have extremely low levels of recombination (Pizza et al.

```
MFKRSVIAMA CIVALSACGG GGGGGSPDVK SADTLSKPAA PVVTEDVGEE
eeeebbbbbb bbbbbbbbbe eeeeeeebee beebeeebbe bbbeeebeeb
ffffssssss ss ssssssf fffffffsff sffsfffssf ss fffsffs 50

VLPKEKKDEE AVSGAPQADT TQQDATAGKG GQDMAAVSAE NTGNGGAATT
beeeeeeeee eeebeeeeee ebeeeeebee bbbbbeeeee eeeeeeeeee
sfffff  ff    sfff  f  s    f f  sssss ffff ffff f fff 100

DNPENKDEGP QNDMPQNAAD TDSSTPNHTP APNMPTRDMG NQAPDAGESA
eeeeeeeeee eeeebeeeee eeeeeeeeee bbbeeeeeeb eeeeeeeeee
 f ff  f f fff sf fff  ffffff ff s    f ffs  f fff fff 150

QPANQPDMAN AADGMQGDDP SAGEENAGNT ADQAANQAEN NQVGGSQNPA
ebeebbebbe ebeeeeeeee eebeeebeeb eeeeeeeeee eeeeeeeeee
fsffssf sf f ffffff f  fsffs f   fff fff   ff       f 200

SSTNPNATNG GSDFGRINVA NGIKLDSGSE NVTLTHCKDK VCDRDDFLDE
eeeeeeeeeb bebebbebbe bebeeeebbb eebeeeeeee ebbeeeeeee
    f f  s sf             f f ss ffsf   f     sfff   f 250

EAPPKSEFEK LSDEEKINKY KKDEQRELEN NNFVGLVADR VEKNGTNKYV
eebeebeeee ebeebeeeee eeebbebbbe ebeeeeeeeb beebeeeeee
ffs  sf        sf        s   ss       f f   s    fff   300

IIYKDKSASS SSARFRRSAR SRRSLPAEMP LIPVNQADTL IVDGEAVSLT
eebbebeebb ebeeebeeeb ebbebeebeb bbbebebbbb bbeebebbbe
   s sff s   fffs  fs fssfsffsfs sssfsfssss ssffsfsss  350

GHSGNIFAPE GNYRYLTYGA EKLSGGSYAL SVQGEPAKGE MLAGTAVYNG
eeebebbbeb beebeebbbb bebeeeebee ebbbbbbbbe bebbbbebee
fffsfsssfs sffs fssss s sfffffsff fssss sssf sfssssf ff 400

EVLHFHMENG RPSPSGGRFA AKVDFGSKSV DGIIDSGDDL HMGTQKFKAV
eeeeeeebeb bbebeeebee bebbbbbbee beebeeebeb bbeeeebebe
ff     sfs ssfsfffsff sfsssssff sffs ffsfs  sfffsfsf 450

IDGNGFKGTW TENGGGDVSG RFYGPAGEEV AGKYSYRPTD AEKGGFG
eeeeeeeeee bebebeeeee bbeeeeeeee eeeeeebbbb beeeeee
fffff ffff s sfsfffff ssfffffff fffffssss sffffff
```

Fig. 2. ConSeq predictions of structure for NMB2132 demonstrated on AF226419, using all homologous genes within this family. The sequence of the query protein is displayed on the first row. The second row lists the predicted burial status of the site (i.e., b, buried, versus e, exposed). The third row indicates residues predicted to be structurally and functionally important: s and f, respectively. Sites inferred as evolving under the influence of positive selection are in boldface and underlined: the majority of such sites are in exposed regions of the protein.

2000). Too few sequences can also affect the power of the LRT, it has been suggested that a minimum of six sequences is required for the results of such an ML analysis to be valid (Anisimova et al. 2003). In all seven datasets there was a minimum of 20 sequences; therefore these concerns are not applicable to these results.

Previous vaccine studies have concluded that vaccines targeted against epitopes which consist of negatively selected sites protect more efficiently than those directed against epitopes which contain positively selected sites (Suzuki 2004). Following an examination of the sites that have undergone positive selection within the HIV genome, it has been suggested that negatively selected sites could be used to design a multiepitope vaccine directed against regions of the virus that are unable to mutate and escape immune recognition (de Oliveira et al. 2004). Using this logic we suggest that of the seven *Neisseria* vaccine candidates analyzed here, careful examination needs to be given to those that exhibit evidence of positive selection. Among a suite of seven proteins that are known to be highly expressed, known to be surface exposed, and known to be capable of eliciting a strong bactericidal immune response, two showed evidence of having undergone positive selection. The other five proteins do not display any historical evidence of positive selection for change. The fact that evidence of positive selection was found in some of these proteins is not particularly surprising, given that other membrane proteins have been shown to evolve positively (Andrews and Gojobori 2004; Jiggins et al. 2002; Smith et al. 1995). Intuitively we would expect these proteins to be under the greatest pressure for change in an effort to remain ahead of host immune responses in the host–pathogen "arms race." What are the implications of this kind of analysis? At this current moment in time, all these proteins still elicit a bactericidal immune response. However, if vaccines target epitopes that contain positively selected sites such as those found in NMB0992 and NMB2132, it is feasible that future amino acid altering substitutions at these sites may render such vaccines obsolete. From this analysis, we would suggest that this kind of event is much more likely in those proteins that have historically shown an ability to evolve under positive selection than in those proteins where positive selection has not been a feature. We therefore propose that an analysis of historical adaptive evolution seems to be a sensible precautionary measure prior to the expensive process of developing a vaccine. We suggest that the

remaining five vaccine targets are ideal vaccine candidates due to the absence of positive Darwinian selection, and any vaccines developed against NMB0992 and NMB2132 should avoid targeting the regions we have inferred to be under the influence of positive selection.

Acknowledgments. We wish to thank the comments of two anonymous reviewers. D.F. was supported by a Higher Education Authority grant (Programme for Research in Third Level Institutes) and C.C. by a Health Research Board grant (RP124/2001).

References

Andrews TD, Gojobori T (2004) Strong positive selection and recombination drive the antigenic variation of the PilE protein of the human pathogen Neisseria meningitidis. Genetics 166:25–32

Anisimova M, Bielawski JP, Yang Z (2001) Accuracy and power of the likelihood ratio test in detecting adaptive molecular evolution. Mol Biol Evol 18:1585–1592

Anisimova M, Nielsen R, Yang Z (2003) Effect of recombination on the accuracy of the likelihood method for detecting positive selection at amino acid sites. Genetics 164:1229–1236

Berezin C, Glaser F, Rosenberg J, Paz I, Pupko T, Fariselli P, Casadio R, Ben-Tal N (2004) ConSeq: the identification of functionally and structurally important residues in protein sequences. Bioinformatics 20:1322–1324

Bjune G, Hoiby EA, Gronnesby JK, Arnesen O, Fredriksen JH, Halstensen A, Holten E, Lindbak AK, Nokleby H, Rosenqvist E (1991) Effect of outer membrane vesicle vaccine against group B meningococcal disease in Norway. Lancet 338:1093–1096

Bush RM, Fitch WM, Bender CA, Cox NJ (1999) Positive selection on the H3 hemagglutinin gene of human influenza virus A. Mol Biol Evol 16:1457–1465

Creevey CJ, Mclnerney JO (2002) An algorithm for detecting directional and non-directional positive selection, neutrality and negative selection in protein coding DNA sequences. Gene 300:43–51

de Oliveira T, Salemi M, Gordon M, Vandamme AM, van Rensburg EJ, Engelbrecht S, Coovadia HM, Cassol S (2004) Mapping sites of positive selection and amino acid diversification in the HIV genome: an alternative approach to vaccine design? Genetics 167:1047–1058

Fares MA (2004) SWAPSC: sliding window analysis procedure to detect selective constraints. Bioinformatics:bth303

Fares MA, Moya A, Escarmis C, Baranowski E, Domingo E, Barrio E (2001) Evidence for positive selection in the capsid protein-coding region of the foot-and-mouth disease virus (FMDV) subjected to experimental passage regimens. Mol Biol Evol 18:10–21

Frasch CE (1989) Vaccines for prevention of meningococcal disease. Clin Microbiol Rev 2 Suppl :S134–S138

Goldschneider I, Gotschlich EC, Artenstein MS (1969) Human immunity to the meningococcus. I. The role of humoral antibodies. J Exp Med 129:1307–1326

Gotschlich EC, Liu TY, Artenstein MS (1969) Human immunity to the meningococcus. 3. Preparation and immunochemical properties of the group A, group B, and group C meningococcal polysaccharides. J Exp Med 129:1349–1365

Grandi G (2003) Rational antibacterial vaccine design through genomic technologies. Int J Parasitol 33:615–620

Jiggins FM, Hurst GD (2002) Host-symbiont conflicts: Positive selection on an outer membrane protein of parasitic but not mutualistic Rickettsiaceae. Mol Biol Evol 19:1341–1349

Kinsella RJ, Fitzpatrick DA, Creevey CJ, McInerney JO (2003) Fatty acid biosynthesis in Mycobacterium tuberculosis: lateral gene transfer, adaptive evolution, and gene duplication. Proc Natl Acad Sci USA 100:10320–10325

Li WH (1993) Unbiased estimation of the rates of synonymous and nonsynonymous substitution. J Mol Evol 36:96–99

Martin D, Cadieux N, Hamel J, Brodeur BR (1997) Highly conserved Neisseria meningitidis surface protein confers protection against experimental infection. J Exp Med 185:1173–1183

Maynard-Smith J, Smith NH (1998) Detecting recombination from gene trees. Mol Biol Evol 15:590–599

Naess A, Halstensen A, Nyland H, Pedersen SH, Moller P, Borgmann R, Larsen JL, Haga E (1994) Sequelae one year after meningococcal disease. Acta Neurol Scand 89:139–142

Nassif X (2002) Genomics of Neisseria meningitidis. 91:419–423

Nassif X, Pujol C, Morand P, Eugene E (1999) Interactions of pathogenic Neisseria with host cells. Is it possible to assemble the puzzle? Mol Microbiol 32:1124–1132

Nielsen R, Yang Z (1998) Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. Genetics 148:929–936

Nowak MA, Anderson RM, McLean AR, Wolfs TF, Goudsmit J, May RM (1991) Antigenic diversity thresholds and the development of AIDS. Science 254:963–969

Pizza M, Scarlato V, Masignani V, Giuliani MM, Arico B, Comanducci M, Jennings GT, Baldi L, Bartolini E, Capecchi B, Galeotti CL, Luzzi E, Manetti R, Marchetti E, Mora M, Nuti S, Ratti G, Santini L, Savino S, Scarselli M, Storni E, Zuo P, Broeker M, Hundt E, Knapp B, Blair E, Mason T, Tettelin H, Hood DW, Jeffries AC, Saunders NJ, Granoff DM, Venter JC, Moxon ER, Grandi G, Rappuoli R (2000) Identification of vaccine candidates against serogroup B meningococcus by whole-genome sequencing. Science 287:1816–1820

Posada D, Crandall KA (1998) MODELTEST; testing the model of DNA substitution. Bioinformatics (Oxford, England) 14:817–818

Sierra GV, Campa HC, Varcacel NM, Garcia IL, Izquierdo PL, Sotolongo PF, Casanueva GV, Rico CO, Rodriguez CR, Terry MH (1991) Vaccine against group B Neisseria meningitidis: protection trial and mass vaccination results in Cuba. NIPH Ann 14:195–207, discussion 208–110

Smith NH, Maynard-Smith J, Spratt BG (1995) Sequence evolution of the porB gene of Neisseria gonorrhoeae and Neisseria meningitidis: evidence of positive Darwinian selection. Mol Biol Evol 12:363–370

Suzuki Y (2004) Negative selection on neutralization epitopes of poliovirus surface proteins: implications for prediction of candidate epitopes for immunization. Gene 328:127–133

Suzuki Y, Nei M (2001a) Reliabilities of parsimony-based and Likelihood-based methods for detecting positive selection at single amino acid sites. Mol Biol Evol 18:2179–2185

Suzuki Y, Nei M (2001b) Reliabilities of parsimony-based and likelihood-based methods for detecting positive selection at single amino acid sites. Mol Biol Evol 18:2179–2185

Suzuki Y, Nei M (2004) False positive selection identified by ML-based methods: Examples from the Sig1 gene of the diatom thalassiosira weissflogii and the tax gene of a human T-cell lymphotropic virus. Mol Biol Evol:msh098

Swanson WJ, Yang Z, Wolfner MF, Aquadro CF (2001) Positive Darwinian selection drives the evolution of several female reproductive proteins in mammals. Proc Natl Acad Sci U S A 98:2509–2514

Swofford D (1998)) PAUP*: Phylogenetic analysis using parsimony (*and other methods). Sinauer Associates, Sunderland, MA

Tettelin H, Saunders NJ, Heidelberg J, Jeffries AC, Nelson KE, Eisen JA, Ketchum KA, Hood DW, Peden JF, Dodson RJ, Nelson WC, Gwinn ML, DeBoy R, Peterson JD, Hickey EK, Haft DH, Salzberg SL, White O, Fleischmann RD, Dougherty BA, Mason T, Ciecko A, Parksey DS, Blair E, Cittone H, Clark EB, Cotton MD, Utterback TR, Khouri H, Qin H, Vamathevan J, Gill J, Scarlato V, Masignani V, Pizza M, Grandi G, Sun L, Smith HO, Fraser CM, Moxon ER, Rappuoli R, Venter JC (2000) Complete genome sequence of Neisseria meningitidis serogroup B strain MC58. Science 287:1809–1815

Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res (online) 22:4673–4680

Urwin R, Holmes EC, Fox AJ, Derrick JP, Maiden MC (2002) Phylogenetic evidence for frequent positive selection and recombination in the meningococcal surface antigen PorB. Mol Biol Evol 19:1686–1694

Yang Z (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. Comput Appl Biosci 13:555–556

Yang Z (1998) Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. Mol Biol Evol 15:568–573

Yang Z (2000) Maximum likelihood estimation on large phytogenies and analysis of adaptive evolution in human influenza virus A. J Mol Evol 51:423–432

Yang Z (2001) Maximum likelihood analysis of adaptive evolution in HIV-1 gp120 env gene. Pac Symp Biocomput 2001i: 226–237

Yang Z, Nielsen R, Goldman N, Pedersen AM (2000) Codon-substitution models for heterogeneous selection pressure at amino acid sites. Genetics 155:431–449

Zanotto PM, Kallas EG, de Souza RF, Holmes EC (1999) Genealogical evidence for positive selection in the nef gene of HIV-1. Genetics 153:1077–1089