J Mol Evol (2007) 64:411–422
DOI: 10.1007/s00239-006-0021-2

JOURNAL OF MOLECULAR EVOLUTION

Positively Selected Codons in Immune-Exposed Loops of the Vaccine Candidate OMP-P1 of *Haemophilus influenzae*

Ted H. M. Mes,[1] Jos P. M. van Putten[2]

[1] Netherlands Institute of Ecology (NIOO-KNAW), Centre for Estuarine and Marine Ecology, P.O.B. 140, 4400 AC, Yerseke, The Netherlands
[2] Department of Infectious Diseases and Immunology, Utrecht University, Yalelaan 1, 3584 CL Utrecht, The Netherlands

Received: 27 January 2006 / Accepted: 11 January 2007 [*Reviewing Editor:* Dr. Rasmus Nielsen]

Abstract. The high levels of variation in surface epitopes can be considered as an evolutionary hallmark of immune selection. New computational tools enable analysis of this variation by identifying codons that exhibit high rates of amino acid changes relative to the synonymous substitution rate. In the outer membrane protein P1 of *Haemophilus influenzae*, a vaccine candidate for nontypeable strains, we identified four codons with this attribute in domains that did not correspond to known or assumed B- and T-cell epitopes of OMP-P1. These codons flank hypervariable domains and do not appear to be false positives as judged from parsimony and maximum likelihood analyses. Some closely spaced positively selected codons have been previously considered part of a transmembrane domain, which would render this region unsuited for inclusion in a vaccine. Secondary structure analysis, three-dimensional structural database searches, and homology modeling using *FadL* of *E. coli* as a structural homologue, however, revealed that all positively selected codons are located in or near extracellular looping domains. The spacing and level of diversity of these positively selected and exposed codons in OMP-P1 suggest that vaccine targets based on these and conserved flanking residues may provide broad coverage in *H. influenzae*.

Correspondence to: Ted H. M. Mes; *email:* t.mes@nioo.knaw.nl

Key words: *Haemophilus influenzae* — Outer membrane proteinI — Immune selection — Structural information — Epitope — Vaccine

Introduction

Vaccination with *Haemophilus influenzae* type b polysaccharide (in the form of Hib conjugate vaccines) has proven to be highly effective in preventing invasive *H. influenzae* disease and has effectively reduced serotype b in many parts of the world (Black et al. 1992). Nontypeable *H. influenzae* strains, however, remain a serious problem (Zeckel et al. 1992). These strains are generally less invasive than their encapsulated type b counterpart but are a major cause of ear infections (otitis media) and sinusitis in children. They are also associated with respiratory tract infections such as pneumonia in infants, children, and adults. Ongoing efforts to develop a vaccine against nontypeable *H. influenzae* have mainly focused on immunogenic surface-exposed proteins (Bolduc et al. 2000). This approach clearly has potential, as immune responses against surface proteins have been shown to aid the recovery from otitis media (Shurin et al. 1980) and vaccination with surface-exposed domains of a major outer membrane protein (OMP-P1) provides protection in animal models (Bolduc et al. 2000). However, this effect is strain-specific (Gonzales et al.

1987) and, hence, not sufficient to provide broad protection. This problem may be overcome by the inclusion of less variable immunogenic protein regions in the vaccine.

Identification of candidate vaccine antigens is classically achieved via immunization studies with isolated or recombinant antigen and via epitope mapping, often in combination with analysis of correlates of protection. Vaccine development is often hampered by antigenic diversity within the surface-exposed regions. This limits a broad cross-reactivity of the elicited immune response. This holds also for proteins such as OMP-P1 in which large conserved domains separate relatively small variable domains (Munson and Grass 1988; Chong et al. 1995; Bolduc et al. 2000). On the other hand, the genetic diversity of vaccine candidates can be used to choose OMP-P1 variants for use in immunological assays (Bolduc et al. 2000). Nowadays, instead of an arbitrary selection of vaccine candidates, novel and rigorous computational approaches can be used to identify codons whose diversity is driven by the immune system. These codons are characterized by a higher rate of nonsynonymous substitutions (dN) relative to the synonymous substitution rate (dS) (Yang et al. 2003; Fitzpatrick and McInerney 2005). The codons are referred to as positively selected, although they may typically reflect immune selection (Fitzpatrick and McInerney 2005). To avoid confusion with the results of immunological studies, we refer here to codons with a higher rate of nonsynonymous substitution than synonymous substitution (dN > dS) as positively selected.

Knowledge of the position of positively selected codons in a protein would constitute an excellent starting point for immunization studies and epitope mapping, not only because of their biological function, but also because the number of variable sites of proteins that are candidates for inclusion in a vaccine can be reduced. In the present study, we identified codons that evolved more rapidly through nonsynonymous than through synonymous substitutions in a sample of 36 OMP-P1 sequences. We compared the location of these codons with the location of peptides that were used in epitope mapping and with B- and T-cell OMP-P1-specific antigens to examine the congruence among these techniques and to identify regions that could be important for vaccine design. Finally, we localized stretches with positively selected codons in secondary structures and three-dimensional (3D) models of OMP-P1. Our computational approach led to the identification of several novel domains with positive selected codons within the OMP-P1 protein that may be attractive targets in future vaccine design.

Materials and Methods

Evolutionary Analysis of the Selection Pressure on OMP-P1

We used the nr database for BLASTP searches under standard settings to collect closely related OMP-P1 sequences of *H. influenzae* using accession 9716616 as a query. The GenBank accession numbers and the data set of OMP-P1 sequences are provided as Supplementary Information. We aligned the sequences using the standard alignment parameters using Clustalx (Thompson et al. 1997) and checked the amino acid alignment with DNASP version 4.0 (Rozas et al. 2003). Codons comprised in insertions and deletions and in extremely variable stretches were removed from the alignment and are referred to in this work as hypervariable domains. A phylogenetic tree was reconstructed using PAUP* (Swofford 2003) with the maximum likelihood algorithm (100 random additions of taxa, TBR branch swapping) under the optimal nucleotide substitution model based on modeltest version 3.06 (Posada and Crandall 1998). Bootstrapping based on maximum likelihood was used to assess support for internodes using 100 random additions, SPR branch swapping, steepest descent, chuckscore = 0.1, and nchuck = 1. We used the alignment and tree topology of one of the maximum likelihood (ML) trees as input for PAML (Yang 1997). This program implements the currently most widely used algorithms to detect selection on codons, which take into account the transition–transversion ratio and codon usage bias and allow identification of codons that accumulate nonsynonymous mutations more rapidly than synonymous ones. To detect such sites, the rates of nonsynonymous and synonymous substitutions are determined through ML analyses using the model of Goldman and Yang (1994). For immune-exposed proteins of pathogens, a dN/dS ratio >1 indicates that nonsynonymous substitutions are driven by selection pressures exerted by the immune system (Derrick et al. 1999; Jiggins et al. 2002; Fitzpatrick and McInerney 2005). Because it is most likely that only a subset of codons in any protein-coding gene evolves under positive selection, several distributions of dN/dS that reflect neutral (M1, M7) and selection (M2, M8) models can be fitted to the data using PAML (Yang 1997). The neutral models assume a single estimate of ω ($0<\omega<1$; M1) and a discrete beta distribution of ω with 10 categories ($0<\omega<1$; M7). The dN/dS ratio under the neutral models is constrained to values ≤ 1 (Yang et al. 2003; Yang and Swanson 2002). The selection models M2 and M8 have an additional class of ω values that exceed unity. This latter codon category is believed to comprise codons that code for amino acids that are affected by the immune system (Yang et al. 2000) and that, consequently, evolve rapidly at the amino acid level. The cross-category ω's under M7 and M8 were estimated by averaging the ω estimates of individual ω classes weighted by their frequency. We used multiple starting values for the selection models M2 and M8, because these models may have multiple local optima (e.g., Wong et al. 2004). The fit of M1 and M7 to the data was compared with selection models (M2 and M8, respectively). The ratio of the fit of neutral and selection models mentioned above follows a chi-square distribution with two degrees of freedom, i.e., for significance at the 1% significance level, two times the difference of the likelihood of neutral and selection models should be >9.21. We examined the evidence for positive selection in the entire alignment (excluding codons with ambiguity codes and gaps; see above), in a data set comprising 30 nontypeable isolates and in a data set of 36 sequences from which five variable domains were deleted. The latter data set enables the detection of positively selected codons in highly conserved domains. The selection models may give false-positive results in the case of a poor fit of the beta distribution to the true distribution of ω in the interval between 0 and 1 (Swanson et al. 2003). The test advocated by Swanson et al. (2003) to guard against

such false positives of the ML methods implemented in PAML involves the comparison of M8 ($\omega_2 > 1$) and M8A ($\omega_2 = 1$). This likelihood ratio test (LRT) is asymptotically distributed as a 50:50 mixture of a point mass at 0 and a chi-square with one degree of freedom, which corresponds to 2.71 units difference between the likelihoods of M8A and M8 at the 1% level. If the fit of the selection models is significantly better than that of the corresponding neutral models, the selection models can be used to identify individual codons under positive selection (Yang et al. 2005). For inferring which sites are under positive selection, the Bayes theorem was used (Yang et al. 2005), which assumes that given the proportion of sites across ω classes as estimated from the ML analyses, assigns priors to model parameters to integrate over their uncertainties. The Bayes Empirical Bayes (BEB) approach used here was specifically advocated for small data sets, because it takes into account sampling errors of the proportions and ω ratios of site classes (Yang et al. 2005).

The rigorous identification of codons under positive selection is an important step when linking evolutionary and structural data, and when testing individual codons, p-values should ideally be confirmed by independent tests and corrected for multiple tests. To these ends, we applied the conservative tests of positive selection based on parsimony (Suzuki and Gojobori 1999) implemented as the Single Likelihood Ancestor Counting (SLAC) analysis in the datamonkey web server (Kosakovsky Pond and Frost 2005). In contrast to the original version of the Suzuki-Gojobori test, the modified version allows selection of a nucleotide substitution model. We also used a new ML method developed by Massingham and Goldman (2005), which does not assume an underlying distribution of ω. Instead, their sitewise likelihood ratio (SLR) procedure, which uses the same codon substitution model as in Goldman and Yang (1994), performs a LRT on a sitewise basis, testing the null model (neutrality, $\omega = 1$) against an alternative model ($\omega \neq 1$). In the SLR method, tree topology, branch lengths, equilibrium codon frequencies, and transition/transversion rate ratio are assumed to be common to all sites in an alignment. Under the null model, all parameters except ω are allowed to vary freely. Under the alternative model, all parameters vary freely. The SLR method applies Hochberg's step-up procedure to correct for multiple tests (Massingham and Goldman 2005). Both the parsimony method and the SLR method also allow the identification of negatively selected sites.

Secondary Structure of OMP-P1

Following the localization of positively selected codons in the primary sequence, we determined the higher-order structure of amino acid stretches in OMP-P1 using TMHMM version 2.0 (Krogh et al. 2001) to determine whether stretches with positively selected codons take secondary structures typical of transmembrane domains. The exposition of regions of OMP-P1 was assessed by means of the solvent accessibility algorithm as implemented in Jpred (Cuff et al. 1998). We identified exposed residues using the secondary structure algorithm implemented in SSPRO 4.5 (Cheng et al. 2005) and Jnet5 (Cuff and Barton 1999). These attributes may suggest whether regions with positively selected codons are exposed to the immune system.

Tertiary Structure of OMP-P1

We examined the 3D structure of OMP-P1 using x-ray crystallography or NMR data on structurally related proteins. We used the precompiled Vector Alignment Search Tool (VAST) and NCBI's structural database to identify closely related structural variants that could serve as a template for homology modeling of OMP-P1. The entries in VAST contain experimentally verified information on secondary structure elements such as type, relative orientation, and connectivity. Structurally related proteins can be fitted to each other on a residue-by-residue basis using VAST's transformation matrices and a Gibbs sampling algorithm. In addition, alternative alignments can be generated which take into account the information in secondary structure, thereby increasing the confidence one can have in amino acid alignments of relatively divergent proteins. The VAST program was used for one randomly chosen *H. influenzae* OMP-P1 sequence (GenBank accession no. 9716616 of the nontypeable strain 1512A). Superimposed models of template and target protein were constructed to illustrate the position of gaps and insertions, hypervariable domains, and positively selected codons.

As an alternative to the identification of structural neighbors as implemented in VAST, homology modeling can be used. This always starts by looking for close relatives using standard PSI-BLAST or other similarity searches. We used homology modeling as implemented in GENO3D (Combet et al. 2002) and SWISS-MODEL (Peitsch 1996; Schwede et al. 2003) to assess the congruence with the VAST searches and to collect additional data on liability of portions of the 3D models of OMP-P1 and its structurally closest neighbors. The 3D models of OMP-P1 were visualized using Cn3d version 4.1, a free 3D model viewing and selection program. To judge the quality of the 3D reconstruction, we collected the percentage identity between template and target, the VAST P-value, the score reported by this program, and the squared root of mean square deviations (RMSD) of the backbone atoms between the template and the target proteins in angstroms (Å) using a sliding window approach. To assess whether hypervariable domains were robustly assigned to their exact coordinates in the 3D structures, we also examined the reports of loop reconstructions by the CSP (Constrained Space Programming) routine as implemented in Swiss-Model (Schwede et al. 2003).

To further quantify the degree of similarity in β-strands between OMP-P1 and a structural template, we compared the distribution of amino acids maintaining β-barrels of transmembrane proteins using the TBB-PRED program (Liu et al. 2003), which uses atomic data to assign the most likely amino acid stretches that contribute to β-barrels. We used the Support Vector Machines (SVM) algorithm to determine the location of β-barrels.

Results

Alignment of OMP-P1 Sequences

BLASTP searches used cutoffs of 0 (E-value) and 727 (score) to identify closely related OMP-P1 sequences. Thirty-six hits were obtained, including 33 protein-coding sequences (27 nontypeable, 1 a, 3 b, 1 e, and 1 f) that have previously been analyzed as part of a population genetics study of OMP-P1 (Bolduc et al. 2000; six sequences were left out because they were not among the top hits of the BLASTP search). The average pairwise amino acid similarity in the entire data set was 94%. The sequences of the nontypeable and typeable OMP-P1 protein variants comprised three hypervariable portions with many insertions and deletions (nucleotides 262–282, 610–621, plus 625–633 and 1288–1302 plus 1309–1317) and a few codons with ambiguity codes or gaps (646–651, 730–732, 850–852, 877–879, and 1360–1362). In addition, 22 codons missed from the 5' end of the OMP-P1 gene of isolate 88591, leading to a total of 50 codons

that were deleted to ensure a correct alignment. The amino acid and the nucleotide alignments of the five most variable domains that remained after removal of the codons mentioned above can be viewed as Supplementary Information. The nucleotide alignment revealed that many of the underlying codons in the variable domain frequently shared nucleotides at two of three positions, indicating that the alignment of these regions was correct.

Identification of Positively Selected Codons

The computational approaches implemented in PAML version 3.14 (Yang 1997) determine which, if any, codons may have evolved under positive selection. On the basis of a single phylogenetic estimate using the preferred nucleotide substitution model derived from MODELTEST (the transversion model TVM, which has variable base frequencies, variable transversion rates, and equal transition rates) with a proportion of invariant sites (0.66) and gamma distributed substitution rates (shape parameter, 0.51), we found 14 ML trees, one of which is shown in Fig. 1 (ML –5104.27). Bootstrapping indicated that four branches received 100% support (Fig. 1). Support for the other branches was $<95\%$ (not shown in Fig. 1). Comparison of neutral and selection models using the OMP-P1 data set and the tree in Fig. 1 indicated that both selection models M2 (likelihood, –5169.38) and M8 (–5177.30) fitted the data significantly better than the neutral models M1 (–5260.86) and M7 (–5270.13), respectively (Table 1). Different starting values of ω did not lead to different parameter estimates after optimization (not shown). Both comparisons M1–M2 and M7–M8 exceeded by far the difference required for significance at the 1% level (see Materials and Methods). This suggested that some codons had a higher nonsynonymous than synonymous substitution rate. Similarly, the Swanson test indicated significant evidence for the occurrence of positive selection under M8 (Table 1). Both selection models M2 and M8 suggested that the same codons were under positive selection, i.e., codons 93, 94, 97, 99, 105, 198, 222, 284, and 329, and possibly also codon 96 (under model M8). The OMP-P1 alignment (Supplementary Information) further showed that in this sample of OMP-P1 sequences, positively selected codons mostly flank hypervariable domains. Limitation of the analyses to the 30 nontypeable strains resulted in very similar findings. In these analyses, both M2 (likelihood, –4802.93) and M8 (–4808.87) fitted the data significantly better than M1 (–4883.86) and M7 (–4895.19), respectively (not shown). Thus, both model comparisons were again significant. This held also true for the positively selected codons. For M2, relative to the analyses in Table 1, one additional positively selected codon (225) was found. For M8, no additional codons were found. After deletion of the five variable domains, no evidence for positive selection was left (M1, –3107.88; M2, –3107.77; M7, –3108.18; M8, –3106.35).

The modified Suzuki-Gojobori method indicated that four sites evolved under positive selection (substitution model 012010; codons 93 [$P=0.01$], 94 [$P=0.01$], 97 [$P=0.03$], and 105 [$P=0.03$]; likelihood, –5548.10; tree length per site, 0.59, dN/dS over all codons = 0.32; middle row in Fig. 2). Using the SLR method, we found eight positively selected codons. Relative to the results of PAML (top row in Fig. 2), all but one positively selected codon found by the SLR method remained significant upon correction for multiple tests (except for codon 329, $P=0.14$; bottom row in Fig. 2 and Table 1). Overall, the strongest evidence for positive selection based on the parsimony and ML methods was found for residues 93, 94, 97, and 105.

Topographic Localization of the Amino Acids Encoded by Putatively Positively Selected Codons and Their Relation to Immunological Data of OMP-P1

A priori, positively selected codons in B-cell and T-helper determinants in the OMP-P1 protein are suitable vaccination targets, because these sites can be expected to be highly immunogenic and involved in the clearance of OMP-P1 variants. The most promising broadly protective regions of OMP-P1 based on epitope mapping studies are mainly located outside the stretch of closely spaced positively selected codons (Fig. 3). Peptide-specific antisera directed against these regions did not invoke protective immunity, although strong IgG and T-cell responses were obtained (Proulx et al. 1991; Chong et al. 1995). Only in rare cases was targeting of the conserved regions associated with protection (one epitope located on peptide HIBP1-4 and recognized by MAb 7C8 was protective [Panezutti et al. 1993]). Furthermore, two of five positively selected sites are present in a synthetic P1 peptide (HIBP1-12) that has been used in immunization studies (Proulx et al. 1992; Panezutti et al. 1993; Chong et al. 1995). The peptide, which was not identified as a conserved B-cell epitope across typeable and nontypeable *H. influenzae* isolates, did not carry an immunodominant epitope in animal models. Another peptide that was used for epitope mapping and immunological assays (Proulx et al. 1992; Chong et al. 1995) fully overlaps with the stretch with positively selected amino acids. This HIBP1-2 (Chong et al. 1995) or 2H (Proulx et al. 1992) peptide covers amino acids 60–88 of 1H in *H. influenzae* serotype b strain MinnA (Munson and Grass 1988), which correspond to residues 82–108 in

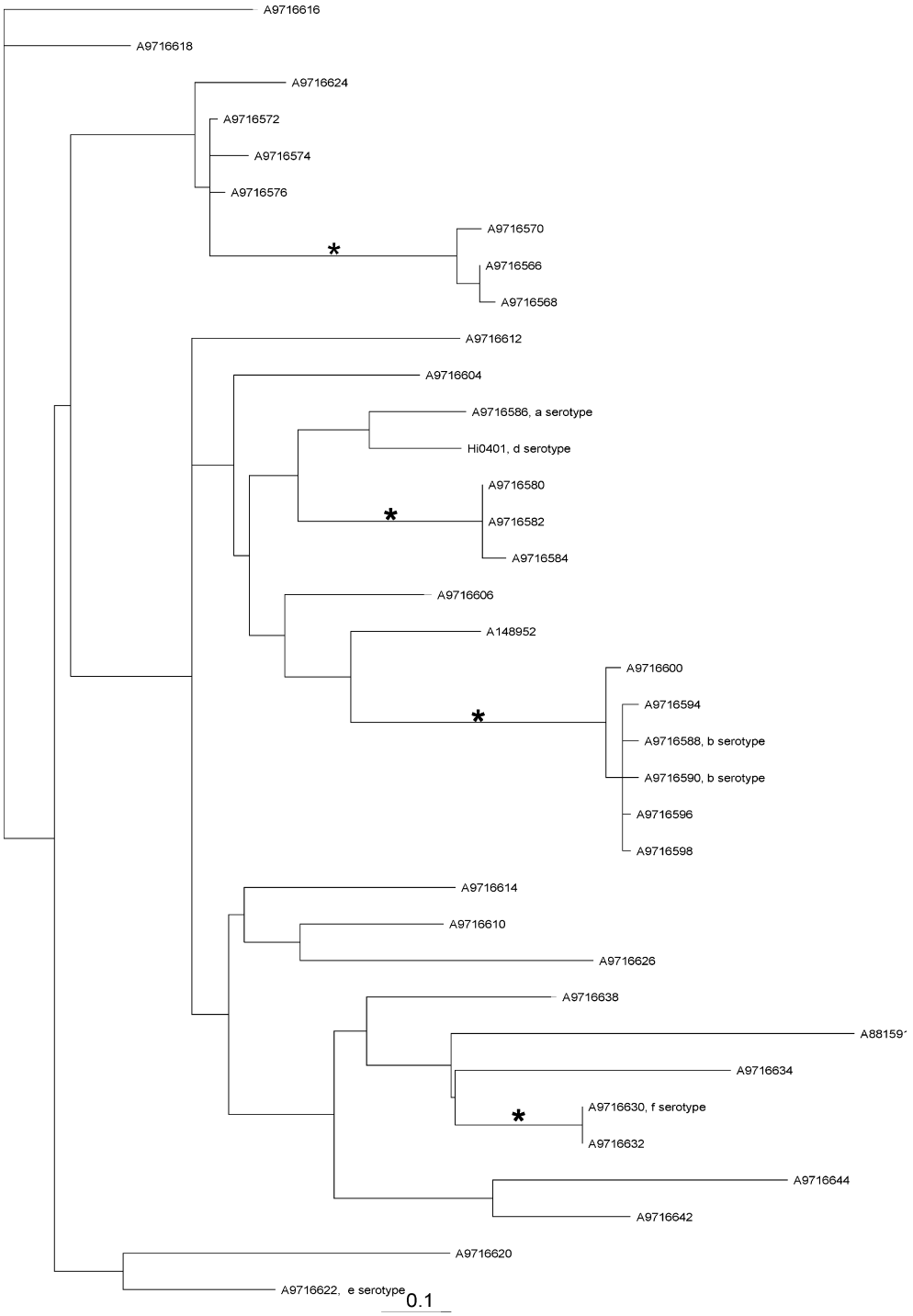

Fig. 1. Maximum likelihood tree based on the TVM model of nucleotide substitution with a proportion of invariant sites and rate heterogeneity (see text). The error bar indicates the number of substitutions per nucleotide. Asterisks indicate branches with 100% bootstrap support.

the reference sequence 9716616. However, anti-P1 antisera isolated from rabbits, guinea pigs, and humans did not react with HIBP1-2 (Chong et al. 1995), and vice versa, anti-HIBP1-2 specific sera did not recognize the intact P1 protein. Despite these negative immunological data, our computational analysis suggests that the stretches with positively selected codons are likely under immune selection in vivo. It may be worthwhile to study these stretches in more detail at the immunological level.

Five of the nine codons for which significant evidence for positive selection was obtained in both PAML selection models M2 and M8 cluster tightly in the primary sequence of OMP-P1 (amino acids 93–105) and four of these are confirmed as positively selected sites using the other types of analyses of positive selection (Fig. 2). Chong et al. (1995) found that some of these codons were part of a transmembrane domain based on secondary structure analysis, hydrophobicity, and reactivity with monoclonal

Table 1. Positively selected codons in OMP P1 *Heamophilus influenza* according to neutral models (M1 and M7) and selection models (M2 and M8)

	Likelihood	Tree length	Kappa κ	Parameters	Codon	Amino acid	Probability positive selection (BEB)	dN/dS codon ± SE
M1	−5260.86	1.91	1.90	$\omega_0 = 0.02$, $\omega_1 = 1$, $p_0 = 0.84$, $p_1 = 0.16$			—	—
M2	−5169.38	2.17	2.13	$\omega_0 = 0.02$, $\omega_1 = 1$, $\omega_2 = 6.42$, $p_0 = 0.83$, $p_1 = 0.14$, $p_2 = 0.03$	93	A	1.00	6.523 ± −0.833
					94	S	1.00	6.523 ± 0.833
					96	K	0.57	3.979 ± 2.632
					97	I	1.00	6.523 ± 0.833
					99	R	1.00	6.523 ± 0.834
					105	Q	1.00	6.523 ± 0.833
					198	A	1.00	6.522 ± 0.835
					222	K	1.00	6.523 ± 0.834
					225	T	0.79	5.293 ± 2.349
					284	K	1.00	6.519 ± 0.844
					329	H	1.00	6.493 ± 0.919
M7	−5270.13	2.07	1.91	B(p = 2.14, q = 0.04)			—	—
M8A	−5260.96	1.92	1.91	B(p = 2.14, q = 99), $p_0 = 0.84$, $p_1 = 0.16$, $\omega_2 = 1$				
M8	−5177.30	2.17	2.12	B(p = 0.07, q = 0.38), $p_0 = 0.97$, $p_1 = 0.38$, $\omega_2 = 5.79$	93	A	1.00	5.445 ± 0.585
					94	S	1.00	5.445 ± 0.585
					95	V	0.84	4.661 ± 1.726
					96	K	0.95	5.214 ± 1.099
					97	I	1.00	5.445 ± 0.585
					99	R	0.74	4.191 ±2.015
					100	N	1.00	5.445 ± 0.585
					105	Q	1.00	5.445 ± 0.585
					151	I	0.90	4.929 ± 1.480
					198	A	1.00	5.445 ± 0.585
					222	K	1.00	5.445 ± 0.585
					225	T	0.96	5.238 ±1.098
					284	K	1.00	5.445 ± 0.586
					329	H	1.00	5.442 ± 0.596
					371	Y	0.76	4.279 ± 2.030

The likelihood indicates the fit of the data to the models of Yang (1997). Tree length is measured as the number of mutations per codon. Kappa is the transversion – transition ratio. p_i denotes the proportion of sites falling in site class ω_i. The parameters p and q are the shape parameters of the beta distribution which underlies M8. The probability that codons were under positive selection was determined using Bayes Empirical Bayes (Yang et al. 2005), with the proportion, ω, and standard error indicated per codon. Model 8A refers to the model of Swanson et al. (2003), in which ω_2 is fixed at one under M8. The reference sequence is 9716616.

antibodies. The nature of the variability of the first hypervariable domain has also been questioned on the basis of a lack of hydrophilic residues (Munson et al. 1992). TMHMM version 2.0 was used to determine whether stretches with positively selected codons involved a transmembrane domain or adopted secondary structures typical of such domains (Krogh et al. 2001). No secondary structures such as α-helices and β-sheets typically associated with transmembrane domains were found in this stretch (not shown). Investigation of the exposition of this region using the solvent accessibility algorithm as implemented in Jpred (Cuff et al. 1998) classified this portion of OMP-P1 as an exposed stretch with high solvent accessibility (Fig. 3), which exceeded 25% for the entire stretch with the five positively selected codons. Thus, the positively selected codons are likely to be exposed. The above findings deviate from previous suggestions that these amino acids are part of a putative transmembrane domain (Chong et al. 1995), which would render this stretch unsuitable for vaccine purposes.

Localization of the Amino Acids Encoded by Putative Positively Selected Codons as Assessed by 3D Modeling

The apparent ambiguity about the putative transmembrane domain of OMP-P1 led us to examine the 3D topology of the protein. BLASTP searches identified one 3D structure with high similarity to OMP-P1 of *H. influenzae*; the *FadL* outer membrane

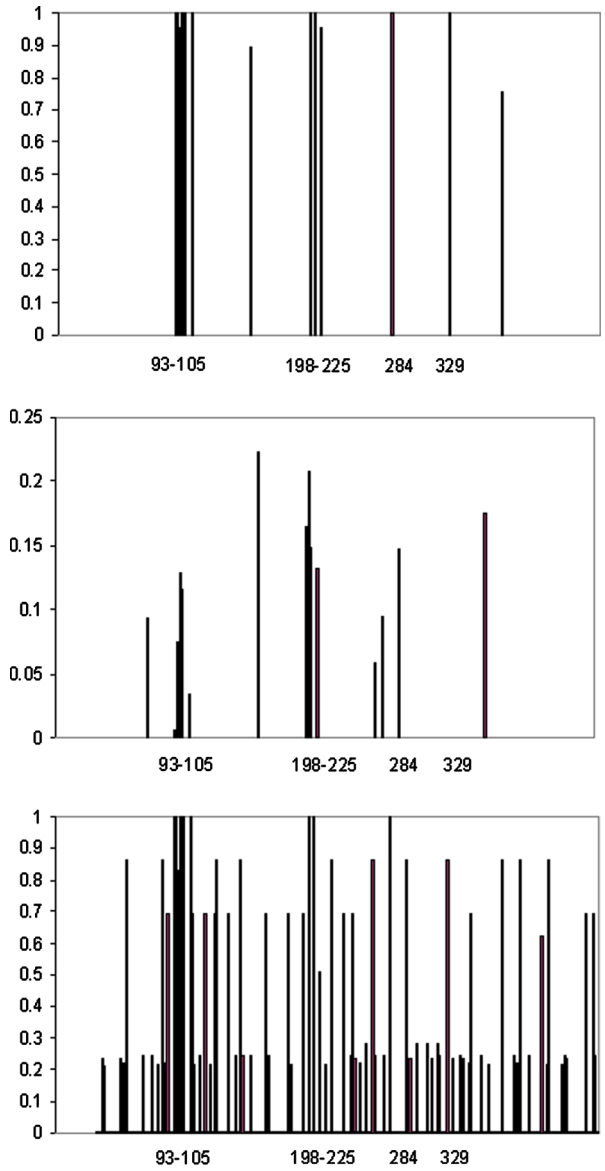

Fig. 2. Codons under positive selection as inferred from the site models in PAML (top panel), the modified Suzuki-Gojobori method implemented as the SLAC procedure in the datamonkey server (middle panel), and the sitewise likelihood ratio test (bottom panel) of 36 OMP-P1 sequences of *H. influenzae*. Codons with significant evidence for positive selection have probabilities >0.95 (PAML and SLR methods) or <0.05 (modified Suzuki-Gojobori method). The spacing between codons is adjusted for the insertions and deletions in OMP-P1.

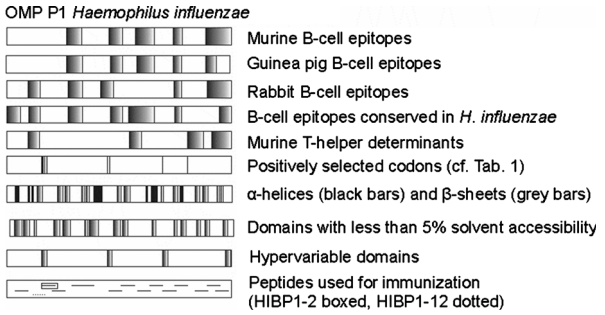

Fig. 3. Diagram of different immunological and genetic attributes of OMP-P1 of *Haemophilus influenzae*. From top to bottom: linear B-cell epitopes in different hosts (murine, guinea pig, and rabbit), B-cell epitopes conserved among *H. influenzae* strains, murine T-helper determinants, positively selected codons, distribution of amino acids with a potentially extended (black) or a helical configuration (gray; Jpred) typical of α-helices and β-sheets, respectively, and distribution of amino acids with $<5\%$ solvent accessibility (Jnet5), location of hypervariable domains, and location of 13 synthetic peptides against OMP-P1, which were used for immunological studies (Proulx et al. 1992; Chong et al. 1995). The approximate locations of peptides HIBP1-2 and HIBP1-12 are boxed and dotted, respectively (see text).

protein of *E. coli* (39% similar in primary amino acid sequence). This protein has been analyzed at a resolution of 2.60 and 2.80 Å using two x-ray crystallography methods (van den Berg et al. 2004). Apart from its role in the transport of long-chain fatty acids across the outer membrane, *FadL* is also a receptor for the bacteriophage T2. OMP-P1 and *FadL* are both classified in PFAM03349.10, which also includes *TodX* from *Pseudomonas putida* F1 and *TbuX* from *Ralstonia pickettii* PKO1. The latter are membrane proteins of uncertain function that are involved in toluene catabolism. The VAST database identified the same structural neighbors (1T1A-B, 1T16A-B) as the BLASTP search and the homology modeling programs GENO3D and SWISS-MODEL (not shown).

There are many indications that the reconstructed model of OMP-P1 (Fig. 4) is reliable. The percentage identity of the OMP-P1 sequences to the target based on secondary elements was approximately 43%, which, in principle, may allow confident 3D models as evidenced by the lower boundary for which homology modeling is used (25%; www.expasy.org/swissmod/SWISS-MODEL.html). Although VAST provides no easily interpretable confidence parameters (the score of this structure pair was 48.5 using an alignment length of 391 residues and the VAST P-value was 10^{-43}), the level of similarity and the root-mean-square deviation (RSMD) of the fit across OMP-P1 and 1T16A indicate a high degree of structural similarity (RMSD $<$ 3 Å; Fig. 5) (T. Madej, NCBI, personal communication). A large number of secondary structures such as helices, strands, and loops are shared between OMP-P1 and *FadL* as evident from VAST searches (Figs. 4A and B). This interpretation is supported by the congruence among models of secondary structure to assign coiled regions that are exposed as determined through the use of SSPRO version 4.5 (not shown) and by the agreement of regions with a β-barrel structure between *FadL* and OMP-P1 as judged from the results of the TBB-PRED server based on atomic data of β-barrel proteins (Fig. 5). Clearly, the fit of the overall structure of the reconstructed 3D model of OMP-P1 does not imply that the model is stable and robust for every region of the protein. The hyper-

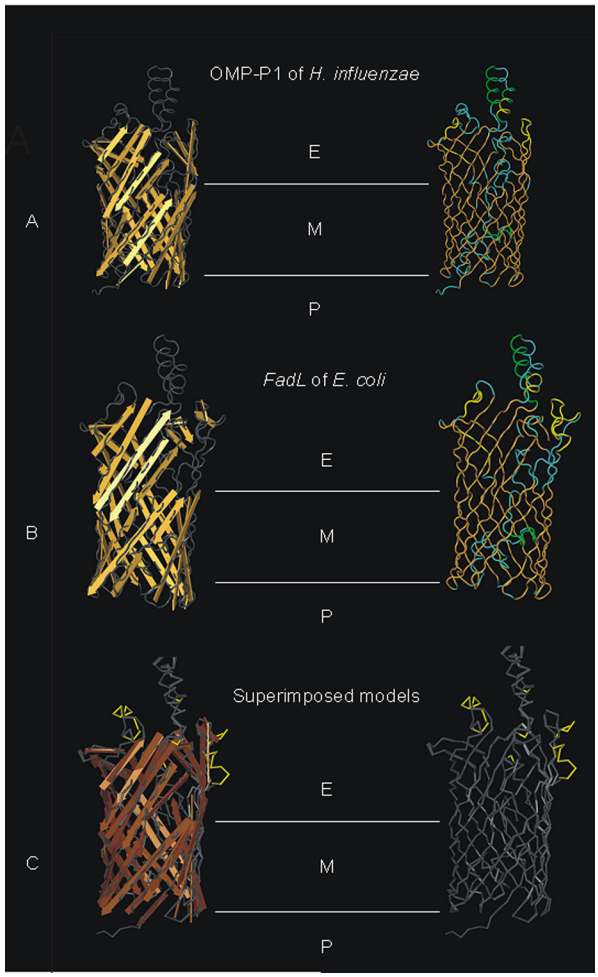

Fig. 4. Three-dimensional model of OMP-P1 of *H. influenzae* (**A**), *FadL* of *E. coli* (**B**), and superimposed models of both proteins (**C**). Brown arrows and threads mark strands, yellow threads mark regions with positively selected codons, green regions mark helical regions, and blue regions mark coiled regions. E, extracellular side; M, membrane; P, periplasm. In the superimposed models, a splitting of the two models indicates gaps or insertions.

variable domains introduce many differences in exposed portions of the 3D models between model and target in the superimposed models (Fig. 4, bottom row). Further, it is important to note that only a single OMP-P1 sequence was used to model the 3D structure based on *FadL*. Slightly different loops may be obtained when using different OMP-P1 sequences. Not surprisingly, the CSP algorithm implemented in SWISS-MODEL had considerable difficulties in finding and optimally position loops (not shown). Many of the loops examined by the program could not be confidently reconstructed, and in many cases, the accepted loops (by the Spare Part procedure) were found only after shorter ones had not been accepted. These regions in many instances corresponded to hypervariable and looping regions (Fig. 4A) that had much higher RMSD values than other regions of the OMP-P1 molecule. However, as judged from the secondary structure analyses these regions likely do not contribute much to the overall stability of the β-barrel structure of OMP-P1 (Fig. 5).

The most relevant conclusion of the 3D reconstruction was that the regions that comprised positively selected codons (yellow in Fig. 4) were always on the extracellular side of the membrane, indicating that the positively selected codons are potential candidates for vaccine development. The positions of insertions and deletions between *FadL* and OMP-P1 that are longer than a few amino acids occurred—in agreement with expectation—mostly in these exposed regions of OMP-P1. Looping regions with insertions and deletions are indicated by the splitting of the two molecules in superimposed 3D models (Fig. 4C). Overall, the superimposed 3D models agree with the models based on individual 3D structures in that, as expected for a β-barrel protein, particularly the main chain hydrogen-bonding patterns characteristic of strands contribute to the structural maintenance of OMP-P1.

Discussion

Positive Selection on OMP-P1

Identification of positive selection on OMPs is generally hampered by extremely high levels of sequence divergence and the frequent occurrence of insertions and deletions in hot spots. These domains are, however, potentially interesting for vaccine studies, as they may be the targets of immune selection (Smith et al. 1995). To enable identification of positively selected codons, we focused on codons directly adjacent to hypervariable and unalignable regions (cf. Supplementary Information). The hypervariable regions were excluded, as it is simply not possible to identify homologous nucleotide positions for these domains. Our strategy to detect positively selected codons is still feasible, as the inference of positive selection is not very sensitive to the use of slightly different trees resulting from, for example, the deletion of hypervariable codons, if sufficient phylogenetic signal remains (Derrick et al. 1999). Our results indeed indicate that after deletion of the hypervariable regions, a substantial number of substitutions remains for OMP-P1 (see Supplementary Information), which allows the identification of positively selected codons.

Another important issue to consider in this type of study is how recombination affects the inferences of positive selection. Recombination generates different tree topologies when based on different regions of a gene, and it has been shown that—depending on the level of recombination—the LRT may be affected (Anisimova et al. 2003). In general, recombination is

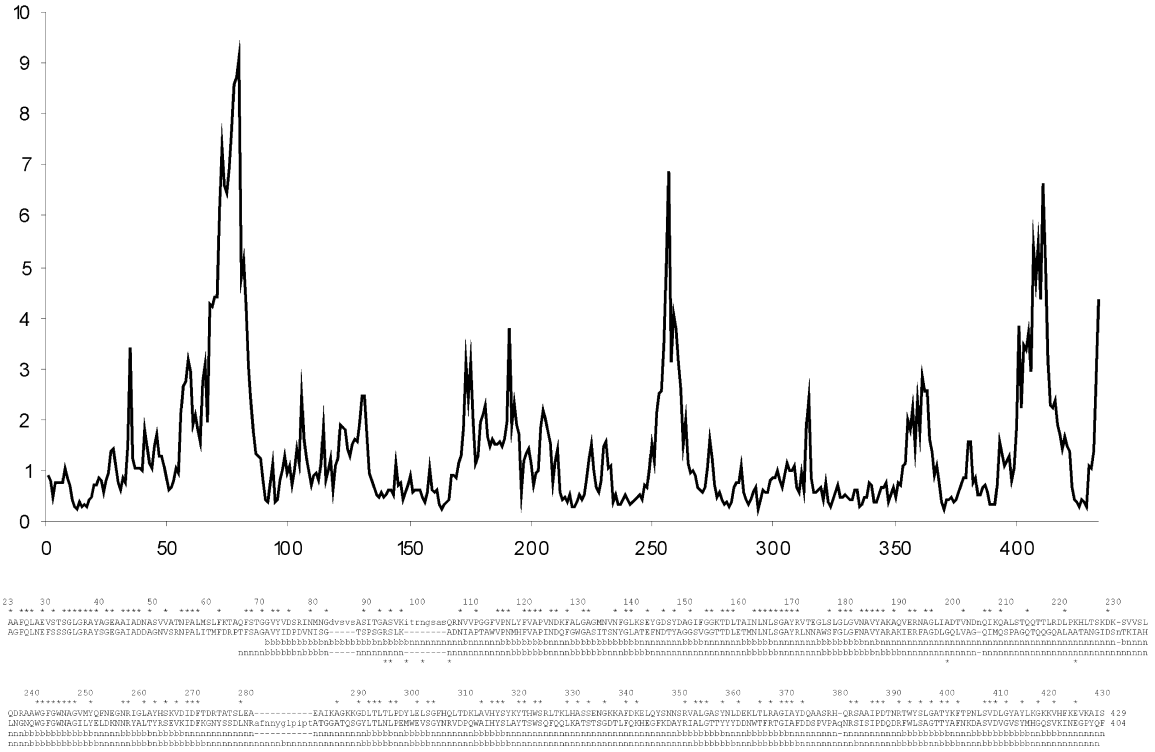

Fig. 5. Alignment and structural analysis of OMP P1 9716616 of *H. influenzae* and *FadL* of *E. coli* based on the VAST 3D structural database. The deviation (Å) of the 3D model of OMP P1 relative to the template *FadL* is shown for windows of five amino acids (above alignment). The contribution of individual residues to the β-barrel structure of OMP P1 as analyzed with TBB-PRED is marked by "b" (barrel) and "n" (nonbarrel). Asterisks above the alignment mark conserved amino acid positions; asterisks below the alignment mark positively selected amino acid positions. The alignment of OMP P1 9716616 starts at residue 23.

considered most problematic for analysis of selection pressures using divergent sequence data (cf. viral sequence data [Anisimova and Yang 2004]), but it has generally been difficult to determine its quantitative importance. Simulation studies provide information on the circumstances in which recombination may affect evolutionary analysis of selection pressures. For example, in spite of the congruent results between M2 and M8 in this study, these models are not equally affected by recombination (Anisimova et al. 2003). In addition, the Bayesian analysis that was used to identify positively selected residues is less affected by the presence of recombination than the LRTs, presumably because it does not depend strongly on the entire gene tree topology but, instead, reconstructs the numbers of synonymous and nonsynonymous mutations of individual codons. Finally, the proportion of correctly classified sites for the BEB procedure in simulations under positive selection and recombination increases with selection pressure (Anisimova et al. 2003). The higher ω values of positively selected codons in this study (Table 1) relative to those of the simulation study mentioned above ($\omega_2 = 2.55$) and the large number of positively selected sites in OMP-P1 suggest that the inference of positive selection and sites under positive selection is valid.

Use and Value of 3D Reconstruction of OMP-P1

The alternation of relatively conserved and hypervariable stretches in OMP-P1 and other OMPs also impacts the 3D reconstruction. Hypervariable amino acid stretches may render a 3D structure unstable, despite the fact that more conserved parts may share considerable similarity between target and template. Thus, although the level of congruence between the experimental 3D structure of *FadL* and the computational 3D structure of OMP-P1 may vary considerably, the most probable and approximate location of stretches with positively selected codons can be determined if the conserved portions of OMP-P1 can be confidently reconstructed. With this goal in mind, the identification of β-barrel regions and the degree of fit of stranded regions between model and template are most important. These attributes leave little doubt that the fit between *FadL* and OMP-P1 is sufficient to pinpoint the approximate localization of positively selected codons (Fig. 5). This is not surprising, as more divergent structural templates and

targets such as *OmpF* and *PhoE* of *E. coli* have successfully been used for the reconstruction of 3D models of porins in *Neisseria* (Derrick et al. 1999). In sum, there is every reason to believe that the basic barrel structure is not profoundly affected by hypervariable loops, even if these experience frequent insertions and deletions that are potentially also under positive selection (Smith et al. 1995; Derrick et al. 1999).

Are Positively Selected Sites Useful for Vaccine Formulations?

The question arises why the peptides that partially (HIBP1-12) and fully (HIBP1-2) overlapped with the stretch of positively selected sites did not include immunodominant epitopes (Proulx et al. 1992; Panezutti et al. 1993; Chong et al. 1995). The stretches, however, were very long (29 and 27 amino acids, respectively) and they extended well into the interior of OMP-P1 (not shown). The potentially most interesting peptide (HIBP1-2) comprises the entire hypervariable domain of the Eagan strain (Chong et al. 1995), including residue 105 (SASQRNV is the C-terminus; Supplementary Information). Our analyses suggest that relatively short peptides starting at the C-terminal end of the first hypervariable domain would be suited for immunological experiments. This portion of OMP-P1 has not yet been explored.

Interestingly, one of the positively selected residues in OMP-P1 that flank the first hypervariable domain allows examination of the impact and prospect of positively selected residues for vaccine formulation in a considerable fraction of the *H. influenzae* population. The positively selected codons 93, 94, and 97 are flanked on one side by the first hypervariable region and are separated from codon 105 by a conserved amino acid stretch (see Supplementary Information). The latter residue, in turn, is flanked by a long region at the C-terminal side that is conserved in the collection of *H. influenzae* strains examined by Bolduc et al. (2000). In our study sample, only three amino acids are encoded by residue 105, and this number increases only marginally in the larger set of *H. influenzae* isolates (Bolduc et al. 2000). Consequently, judging from the spacing of positively selected and conserved residues, there are opportunities to combine sites under purifying and positive selection in immunological experiments of OMP-P1. If it can be demonstrated by, for example, epitope mapping that these positively selected and exposed residues of OMP-P1 are located in immunologically relevant domains, a large portion of the *H. influenzae* population can be targeted by short peptides from this region of OMP-P1. Relative to traditional approaches to vaccine development, this strategy may allow a much broader coverage of the *H. influenzae* population.

Another question is to what extent the isolates studied here reflect the evolutionary diversity of *H. influenzae*. The isolates analyzed by Bolduc et al. (2000) were selected for OMP-P1 sequencing, because they were phylogenetically diverse in a collection of more than 500 *H. influenzae* isolates. Given the increase in multilocus sequence typing (MLST) of human pathogenic bacteria, it is also of interest to compare the diversity of the strains of Bolduc et al. (2000) with the phylogenetic relationships based on housekeeping genes in the MLST database of *H. influenzae*. In the case of *H. influenzae*, seven housekeeping genes are widely used to characterize phylogenetic relationships. In a housekeeping tree of 131 isolates (Meats et al. 2003), most of the nontypeable isolates were comprised of two highly divergent clusters, and in contrast to the housekeeping genes of encapsulated isolates, the congruence among trees based on single housekeeping genes of nontypeable isolates was only marginal. The current view of *H. influenzae* evolution is that nontypeable isolates are genetically distinct from encapsulated variants, and that nontypeable isolates form distinct and diverse populations. Among the five strains that were used by both Bolduc et al. (2000) and Meats et al. (2003), the shared isolates (B-RM7109, d-Rd, a-7416, nt-667, and e-6181) were scattered over the MLST tree. Apart from issues of sampling, the congruence of gene tree topologies based on different housekeeping genes and OMP-P1 hinges on the level of recombination between genes. It is currently unknown how important recombination is in *H. influenzae*. Although information on the clinical symptoms caused by the isolates with OMP-P1 sequences is extremely limited, strains 667 (9716582), BCH-1 (9716566), BCH-2 (9716610), and BCH-3 (9716568) caused middle ear infection, pneumonia, and nasopharyngeal infection in children. In spite of the limited amount of information on the nontypeable isolates with OMP-P1 sequences, the intermingling of nontypeable and typeable isolates on the housekeeping tree and the clinical symptoms of some of the nontypeable isolates suggest that, in principle, the OMP-P1 diversity studied here might be a suited starting point for vaccine development with the potential of providing broad coverage.

Recent technological progress in vaccinology offers hope that epitopes based on moderately rapidly evolving portions of OMPs may turn out to be realistic vaccine material. Although it cannot be excluded that some epitopes will not be stable or immunogenic, or may not constitute stable constructs, extracellular loops in OMPs have been demonstrated to be immunodominant (Easton et al. 2005). Also, the inclusion of multiple OMPs in vaccine formulations is

possible by using outer membrane vesicles or whole-cell vaccines based on recombinant bacteria (de Jonge et al. 2004) and the delivery of antigens based on OMPs of nontypeable *H. influenzae* based on fusion proteins (Riedmann et al. 2003) offer opportunities to target interesting vaccine candidate regions of OMPs. Finally, whole-cell vaccines of LPS mutants with strongly reduced toxicity are increasingly suited for vaccination purposes (Fisseha et al. 2005), thereby reducing the adverse effects of immune responses against these types of vaccines.

Impact of Computational Analyses on Vaccine Development

Positive selection on codons of exposed outer membrane proteins of pathogenic viruses and bacteria is common (Urwin et al. 2002; Yang et al. Yang 2003; Fitzpatrick and McInerney 2005). The lack of correspondence between interesting sites in OMP-P1 of *H. influenzae* in immunological and genetic diversity studies highlights the difficulty of finding the most promising epitopes, even among closely related sequences. The a priori identification of positively selected codons will normally greatly reduce the number of peptides that need to be screened using epitope mapping or immunization. Because the incorporation of antigenic diversity in the experimental setup of immunological and protection studies is a major bottleneck in terms of time and costs, the identification of a limited number of positively selected codons suggests that a multidisciplinary approach comprising evolutionary and structural data and tools can greatly improve the focus of ensuing empirical studies.

Acknowledgments. This is publication 3979 of NIOO-KNAW.

References

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

Anisimova A, Yang Z (2004) Molecular evolution of hepatitis delta virus antigen gene: Recombination or positive selection? J Mol Evol 59:815–826

Black SB, Shinefield HR, Fireman B, Hiatt R (1992) Safety, immunogenicity, and efficacy in infancy of oligosaccharide conjugate *Haemophilus influenzae* Type B vaccine in a United States population—possible implications for optimal use. J Infect Dis 165:S139–S143

Bolduc GR, Bouchet V, Jiang RZ, Geisselsoder J, Truong-Bolduc QC, Rice PA, Pelton SI, Goldstein R (2000) Variability of outer membrane protein P1 and its evaluation as a vaccine candidate against experimental otitis media due to nontypeable *Haemophilus influenzae*: an unambiguous, multifaceted approach. Infect Immun 68:4505–4517

Cheng J, Randall AZ, Sweredoski MJ, Baldi P (2005) SCRATCH: a protein structure and structural feature prediction server. Nucleic Acids Res 33:W72–W76

Chong PL, Yang YP, Persaud D, Haer M, Tripet B, Tam E, Sia C, Klein M (1995) Immunogenicity of synthetic peptides of *Haemophilus influenzae* Type B outer-membrane protein P1. Infect Immun 63:3751–3758

Combet C, Jambon M, Deleage G, Geourjon C (2002) Geno3D: automatic comparative molecular modelling of protein. Bioinformatics 18:213–214

Cuff JA, Barton GJ (1999) Evaluation and improvement of multiple sequence methods for protein secondary structure prediction. Prot Struct Funct Genet 34:508–519

Cuff JA, Clamp ME, Siddiqui AS, Finlay M, Barton GJ (1998) JPred: a consensus secondary structure prediction server. Bioinformatics 14:892–893

de Jonge MI, Hamstra HJ, Jiskoot W, Roholl P, Williams NA, Dankert J, van Alphen L, van der Ley P (2004) Intranasal immunisation of mice with liposomes containing recombinant meningococcal *OpaB* and *OpaJ* proteins. Vaccine 22:4021–4028

Derrick JP, Urwin R, Suker J, Feavers IM, Maiden MCJ (1999) Structural and evolutionary inference from molecular variation in *Neisseria* porins. Infect Immun 67:2406–2413

Easton DM, Smith A, Gallego SG, Foxwell AR, Cripps AW, Kyd JM (2005) Characterization of a novel porin protein from *Moraxella catarrhalis* and identification of an immunodominant surface loop. J Bacteriol 187:6528–6535

Fisseha M, Chen P, Brandt B, Kijek T, Moran E, Zollinger W (2005) Characterization of native outer membrane vesicles from *lpxL* mutant strains of *Neisseria meningitidis* for use in parenteral vaccination. Infect Immun 73:4070–4080

Fitzpatrick DA, McInerney JO (2005) Evidence of positive Darwinian selection in *Omp85*, a highly conserved bacterial outer membrane protein essential for cell viability. J Mol Evol 60:268–273

Goldman N, Yang ZH (1994) Codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol Biol Evol 11:725–736

Gonzales FR, Leachman S, Norgard MV, Radolf JD, Mccracken GH, Evans C, Hansen EJ (1987) Cloning and expression in *Escherichia coli* of the Gene encoding the heat-modifiable major outer-membrane protein of *Haemophilus influenzae* Type B. Infect Immun 55:2993–3000

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

Kosakovsky Pond SL, Frost SDW (2005) Datamonkey: rapid detection of selective pressure on individual sites of codon alignments. Bioinformatics 21:2531–2533

Krogh A, Larsson B, von Heijne G, Sonnhammer ELL (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J Mol Biol 305:567–580

Liu Q, Zhu YS, Wang BH, Li YH (2003) HMM-based method to predict the transmembrane regions of beta-barrel membrane proteins. Comput Biol Chem 27:69–76

Massingham T, Goldman N (2005) Detecting amino acid sites under positive selection and purifying selection. Genetics 169:1753–1762

Meats E, Feil EJ, Stringer S, Cody AJ, Goldstein R, Kroll JS, Popovic T, Spratt BG (2003) Characterization of encapsulated and noncapsulated *Haemophilus influenzae* and determination of phylogenetic relationships by multilocus sequence typing. J Clin Microbiol 41:1623–1636

Munson R, Grass S (1988) Purification, cloning, and sequence of outer-membrane protein P1 of *Hemophilus influenzae* Type B. Infect Immun 56:2235–2242

Munson R, Brodeur B, Chong P, Grass S, Martin D, Proulx C (1992) Outer-membrane proteins P1 and P2 of *Haemophilus influenzae* Type B—Structure and identification of surface-exposed epitopes. J Infect Dis 165:S86–S89

Panezutti H, James O, Hansen EJ, Choi Y, Harkness RE, Klein MH, Chong P (1993) Identification of surface-exposed B-cell epitopes recognized by *Haemophilus influenzae* Type B P1-specific monoclonal antibodies. Infect Immun 61:1867–1872

Peitsch MC (1996) ProMod and Swiss model: Internet-based tools for automated comparative protein modeling. Biochem Soc Trans 24:274–279

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

Proulx C, Munson RS, Grass S, Hamel J, Martin D, Brodeur BR (1991) Identification of a surface-exposed immunodominant epitope on outer-membrane protein P1 of *Haemophilus influenzae* Type B. Infect Immun 59:963–970

Proulx C, Hamel J, Chong P, Martin D, Brodeur BR (1992) Epitope analysis of an immunodominant domain on the P1 protein of *Haemophilus influenzae* Type B using synthetic peptides and antiidiotypic antibodies. Microbial Pathogenesis 12:433–442

Riedmann EM, Kyd JM, Smith AM, Gomez-Gallego S, Jalava K, Cripps AW, Lubitz W (2003) Construction of recombinant S-layer proteins (rSbsA) and their expression in bacterial ghosts—a delivery system for the nontypeable Haemophilus influenzae antigen Omp26. FEMS Immunol Med Microbiol 37:185–192

Rozas J, Sanchez-DelBarrio JC, Messeguer X, Rozas R (2003) DnaSP, DNA polymorphism analyses by the coalescent and other methods. Bioinformatics 19:2496–2497

Schwede T, Kopp J, Guex N, Peitsch MC (2003) SWISS-MODEL: an automated protein homology-modeling server. Nucleic Acids Res 31:3381–3385

Shurin PA, Pelton SI, Tager IB, Kasper DL (1980) Bactericidal antibody and susceptibility to otitis media caused by non-typable strains of *Haemophilus influenzae*. J Pediatr 97:364–369

Smith NH, Smith JM, 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, Gojobori T (1999) A method for detecting positive selection at single amino acid sites. Mol Biol Evol 16:1315–1328

Swofford DL (2003) PAUP*. Phylogenetic analysis using parsimony (*and other methods). Version 4. Sinauer Associates, Sunderland, MA

Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG (1997) The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res 25:4876–4882

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

Urwin R, Russell JE, Thompson EAL, Holmes EC, Feavers IM, Maiden MCJ (2004) Distribution of surface protein variants among hyperinvasive meningococci: implication for vaccine design. Infect Immun 72:5955–5962

van den Berg B, Black PN, Clemons WM, Rapoport TA (2004) Crystal structure of the long-chain fatty acid transporter *FadL*. Science 304:1506–1509

Yang W, Bielawski JP, Yang ZH (2003) Widespread adaptive evolution in the human immunodeficiency virus type 1 genome. J Mol Evol 57:212–221

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

Yang ZH, Swanson WJ (2002) Codon-substitution models to detect adaptive evolution that account for heterogeneous selective pressures among site classes. Mol Biol Evol 19:49–57

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

Yang ZH, Wong WSW, Nielsen R (2005) Bayes empirical Bayes inference of amino acid sites under positive selection. Mol Biol Evol 22:1107–1118

Zeckel ML, Jacobson KD, Guerra FJ, Therasse DG, Farlow D (1992) Loracarbef (Ly163892) versus amoxicillin clavulanate in the treatment of acute bacterial exacerbations of chronic bronchitis. Clin Ther 14:214–229