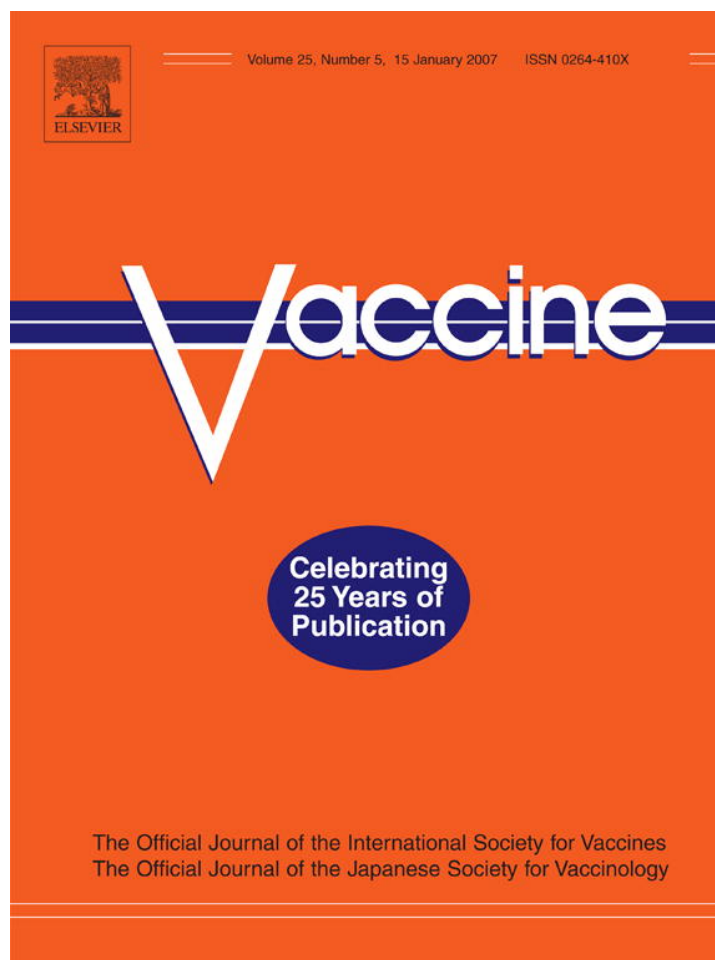


Provided for non-commercial research and educational use only.
Not for reproduction or distribution or commercial use.



This article was originally published in a journal published by Elsevier, and the attached copy is provided by Elsevier for the author's benefit and for the benefit of the author's institution, for non-commercial research and educational use including without limitation use in instruction at your institution, sending it to specific colleagues that you know, and providing a copy to your institution's administrator.

All other uses, reproduction and distribution, including without limitation commercial reprints, selling or licensing copies or access, or posting on open internet sites, your personal or institution's website or repository, are prohibited. For exceptions, permission may be sought for such use through Elsevier's permissions site at:

<http://www.elsevier.com/locate/permissionusematerial>

Identifying candidate subunit vaccines using an alignment-independent method based on principal amino acid properties

Irina A. Doytchinova^{a,*}, Darren R. Flower^b

^a Faculty of Pharmacy, Medical University of Sofia, Dunav st. 2, 1000 Sofia, Bulgaria

^b The Jenner Institute, University of Oxford, Compton, Berkshire RG20 7NN, UK

Received 28 July 2006; accepted 4 September 2006

Available online 28 September 2006

Abstract

Subunit vaccine discovery is an accepted clinical priority. The empirical approach is time- and labor-consuming and can often end in failure. Rational information-driven approaches can overcome these limitations in a fast and efficient manner. However, informatics solutions require reliable algorithms for antigen identification. All known algorithms use sequence similarity to identify antigens. However, antigenicity may be encoded subtly in a sequence and may not be directly identifiable by sequence alignment. We propose a new alignment-independent method for antigen recognition based on the principal chemical properties of protein amino acid sequences. The method is tested by cross-validation on a training set of bacterial antigens and external validation on a test set of known antigens. The prediction *accuracy* is 83% for the cross-validation and 80% for the external test set. Our approach is accurate and robust, and provides a potent tool for the *in silico* discovery of medically relevant subunit vaccines.

© 2006 Elsevier Ltd. All rights reserved.

Keywords: Vaccine; Antigen prediction

1. Introduction

It is now widely accepted that mass vaccination, which takes account of herd immunity, is amongst the most efficacious prophylactic treatments for infectious disease. A vaccine is a molecular or supramolecular agent which elicits specific, protective immunity, which is an enhanced adaptive immune response to re-infection, against pathogenic microbes, and the diseases they cause, by the potentiation of immune memory that ultimately mitigates the effect of subsequent infection. Historically, vaccines have been attenuated whole pathogen vaccines such as BCG for TB or Sabin's Polio vaccine. Issues of safety have led to the development of other strategies for vaccine development, separately focusing on antigen and epitope vaccines. While opinion remains equivocal regarding the ultimate utility of epitope vaccines,

the search for antigen – or subunit – vaccines, such as the Hepatitis B vaccine, is now an accepted clinical priority.

Subunit vaccines contain one or more pure or semi-pure antigens. In order to develop subunit vaccines, it is critical to identify the individual components out of a myriad of proteins and glycoproteins of the pathogen that are involved in inducing protection. Some proteins may be immunosuppressive, whereas in other cases immune responses to some proteins may actually enhance disease. Thus, it is critical to identify those proteins that are important for inducing protection and to eliminate others. The empirical approach to sub-unit vaccine development includes a number of steps beginning with pathogen cultivation, followed by dissection into components, and then testing antigens for their capacity to induce protection [1]. Apart from being time- and labor-consuming, this approach has several limitations that can lead to failure. It only allows the identification of those antigens which can be obtained in sufficient quantities. In some cases, the most abundant proteins are not immunoprotective. In other cases, the antigen expressed during *in vivo* infection is not expressed

* Corresponding author. Tel.: +359 2 9236506; fax: +359 2 9879874.
E-mail address: doytchinova@gmail.com (I.A. Doytchinova).

during *in vitro* cultivation. Vaccines cannot be developed using this approach for non-cultivable microorganisms.

More recently, bioinformatics has been used to select candidate sub-unit vaccine from bacterial and viral genome sequences [2]. These approaches – often termed “reverse vaccinology” [3] – screen the genome sequence *in silico* and predict the most probable protective antigens. They have several obvious advantages: there is no requirement to cultivate pathogens, the whole proteome is available, cost is lower yet speed is high. The bottleneck of these approaches is the development of reliable algorithms for predicting antigens. Many bioinformatics tools have been developed: BLAST [4], FASTA [5], PSORT [6], SignalP [7], GCG [8], which can identify surface-associated or outer membrane proteins, signal proteins, lipoprotein signatures, or host–cell binding domains. Most algorithms use sequence alignment to identify antigens. This is problematic for several reasons. Some proteins created by divergent or convergent evolution lack obvious sequence similarity, although they may share similar structures and biological properties [9]. In such a situation, alignment-based approaches may produce ambiguous results or fail. Moreover, antigenicity, as a property, may be encoded in a sequence in a subtle and recondite manner not amendable to direct identification by sequence alignment. Likewise, the discovery of truly novel antigens will be frustrated by their lack of similarity to antigens of known provenance.

Alternatively, alignment-free methods have been developed where protein sequences are transformed into uniform data matrices. Auto cross covariance (ACC) is an alignment-independent method developed by Wold et al. [10], which has been applied to quantitative structure–activity relationships (QSAR) studies of peptides with different lengths [11,12] and for protein classification [13]. ACC models sequences so that the transformation accounts for neighbor effects, i.e. the lack of independence between different sequence positions. In the present study, we applied ACC preprocessing to a set of known bacterial antigens and developed an alignment-independent model for antigen recognition based on the principal chemical properties of primary amino acid sequences. This represents the first alignment-free bioinformatics tool for the *in silico* identification of antigens.

2. Datasets and methods

2.1. Antigen and non-antigen datasets

The training set consisted of 75 antigens and 75 non-antigens, while the test set included 25 antigens and 25 non-antigens. The antigen subsets comprised known bacterial protein vaccine antigens, taken from the literature [14–52] (Appendix A). A protein was identified as an antigen if it, or part of it, or its corresponding DNA has been shown to induce a protective response in an appropriate animal model after immunization. The non-antigen subsets were constructed to mirror the antigen subsets. For each anti-

gen, a protein was randomly selected from the same species. Proteomes and protein sequences were obtained from the UniProt Knowledgebase of the ExPASy Proteomics Server (<http://ca.expasy.org/sprot/>).

2.2. z -Scales

The z -scales, defined by Hellberg et al. [53], summarize the principal physicochemical properties of the amino acids. These scales were derived by principal component analysis of a data matrix consisting of 29 molecular descriptors, like molecular weight, pK_a s, ^{13}C NMR shifts, etc. The first principle component reflects the hydrophobicity of amino acids, the second their size, and the third their electronic properties. The scores of these components are defined as z_1 -, z_2 - and z_3 -scales, respectively. More recently, Sandberg et al. [54] extended the three z -scales to five, adding two additional z -scales, z_4 and z_5 . By arranging the z -scales according to the amino acid sequence, it is possible to numerically quantify the structural variations within a series of related proteins. In the present study, three z -scales, z_1 , z_2 and z_3 , were used to describe the protein sequences.

2.3. Auto cross covariances (ACC)

As the proteins used in the study had different lengths, an auto cross covariance (ACC) transformation was used to transform them to a uniform length. The auto covariance $A_{jj}(\text{lag})$ was calculated according to Eq. (1) [10]:

$$A_{jj}(l) = \sum_i^{n-l} \frac{Z_{j,i} \times Z_{j,i+1}}{n-l} \quad (1)$$

Index j was used for the z -scales ($j=1-3$), n is the number of amino acids in a sequence, index i the amino acid position ($i=1, 2, \dots, n$) and l is the lag ($l=1, 2, \dots, L$). In order to investigate the influence of close amino acid proximity on protein antigenicity, a short lag of 5 ($L=5$) was used. Cross covariances $C_{jk}(\text{lag})$ between two different z -scales, j and k , were calculated according to Eq. (2) [10]:

$$C_{jk}(l) = \sum_i^{n-l} \frac{Z_{j,i} \times Z_{k,i+1}}{n-l} \quad (2)$$

The results of these transformations were new uniform sets of 45 variables ($3^2 \times 5$) for each protein.

2.4. Variable selection

A genetic algorithm (GA) [55] and stepwise regression, as implemented in the MDL QSAR package [56], were used as variable selection procedures in the present study. GA allows one to select a subset of the most significant predictors using two evolutionary operators: random mutation and genetic recombination (or crossover). Default values for the size of the initial population, choice of parents, types of crossover

and mutation were used in the calculation. Regression equations were generated using variables selected by ordinary multiple linear regression (MLR). Stepwise regression was used in a forward mode using default values for *F*-to-enter (4.00) and *F*-to-remove (3.99). Final models were assessed by ROC statistics.

2.5. Partial least squares (PLS)

PLS forms new variables, named principal components (PC), as linear combinations of the initial variables and then uses them as predictors of the dependent variable. PLS discriminant analysis, as implemented in SIMCA 8.0 [57], was used in the study. The models were assessed by ROC statistics.

2.6. Receiver operating characteristic (ROC) statistics

The correctly predicted antigens and non-antigens were defined as true positives (TP) and true negatives (TN), respectively, while the incorrectly predicted antigens and non-antigens yielded false negatives (FN) and false positives (FP), respectively. Two variables *sensitivity* [$TP/(TP + FN)$] and *1-specificity* [$FP/(TN + FP)$] were calculated at different thresholds and ROC curves were generated [58]. The area under the curve (AUC_{ROC}) is a quantitative measure of the predictive ability and varies from 0.5 for a random prediction to 1.0 for a perfect prediction. Prediction *accuracy* [$(TP + TN)/total$] at different thresholds was also calculated.

3. Results

3.1. Antigen discriminating models

The training set used for the development of antigen discriminating models consisted of 75 known bacterial protein vaccine antigens and 75 non-antigens, selected as described in Section 2. Each amino acid in a protein sequence was described by a set of *z* descriptors: z_1 describes hydrophobicity; z_2 , size; z_3 , electronic properties. As the proteins were of different length, a preprocessing transformation, auto cross covariance (ACC), was used to

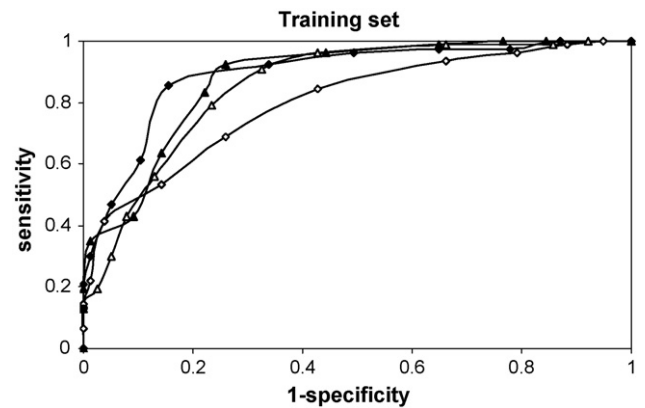


Fig. 1. ROC curves of the models for antigen recognition generated on the training set. The models are marked as follows: PLS (LOO) (▲); PLS 7CV (△); GA (◆); stepwise (◇).

generate a new uniform vector of 45 variables for each protein. The matrix of 45 columns and 150 rows was used to discriminate between antigens and non-antigens using several statistical methods. Models were assessed in terms of AUC_{ROC} , highest *accuracy*, *sensitivity*, and *specificity*. The statistics of the generated models is shown in Table 1 and the ROC curves in Fig. 1. Partial least squares discriminant analysis (DA-PLS) was applied as implemented in SIMCA 8.0. Cross-validation (CV) by “leave-one-out” (LOO) and in seven groups was used to assess the discriminating ability of the PLS model. Both CVs perform well, with an AUC_{ROC} value of 0.876 and 0.853 and highest *accuracy* of 83% and 79%, respectively, at threshold 0.4.

Two variable selection procedures – genetic algorithm (GA) and forward stepwise regression – were applied to the initial uniform matrix to select the best discriminating variables. The GA model gave $AUC_{ROC} = 0.890$ and highest *accuracy* of 85% (threshold 0.5). The stepwise regression was worse, with $AUC_{ROC} = 0.797$ and highest *accuracy* 71% (threshold 0.5).

In terms of *sensitivity* and *specificity*, the PLS models showed higher *sensitivity* than *specificity*, i.e. they are better predictors of antigens than non-antigens. The GA and stepwise models predict antigens and non-antigens almost equally well.

Table 1
Statistics of the antigen discrimination models

Set	Method	AUC_{ROC}	Threshold ^a	Accuracy (%) ^b	Sensitivity (%) ^c	Specificity (%) ^d
Training set (<i>n</i> = 150)	PLS LOO	0.876	0.4	83	92	74
	PLS 7CV	0.853	0.4	79	91	68
	GA	0.890	0.5	85	86	84
	Stepwise	0.797	0.5	71	69	74
Test set (<i>n</i> = 50)	GA	0.756	0.5	72	56	88
	PLS	0.802	0.5	80	64	92

^a The threshold, at which the accuracy is highest.

^b The highest accuracy.

^c Sensitivity at the threshold of the highest accuracy.

^d Specificity at the threshold of the highest accuracy.

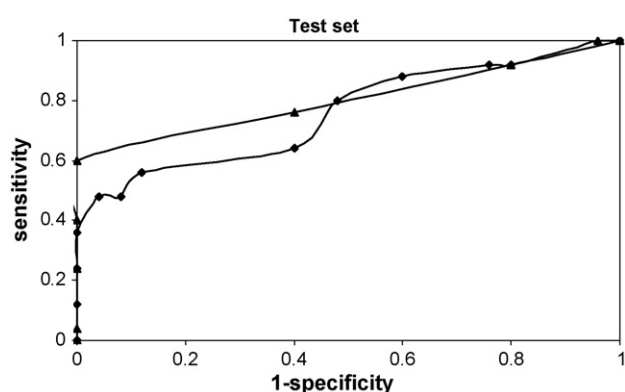


Fig. 2. ROC curves of PLS (LOO) (\blacktriangle) and GA (\blacklozenge) models validated on the test set.

3.2. Validation of the models

The best-performed models PLS (LOO) and GA were applied to an external test set of 25 antigens and 25 non-antigens. The results are shown in Table 1 and the ROC curves plotted in Fig. 2. GA model had a $AUC_{ROC} = 0.756$ and highest $accuracy = 72\%$ (threshold 0.5), while the PLS model gave $AUC_{ROC} = 0.802$ and highest $accuracy = 80\%$ at the same threshold. Both models are better predictors of non-antigens than antigens.

3.3. Variable importance

PLS (LOO) and GA models were compared in terms of variable importance. The higher the variable coefficient the more important it is for accurately discriminating between classes. The sign of the coefficient is arbitrary. In this study, positives contribute positively to the antigen class and negatively to the non-antigen class. The opposite is true for the negative coefficients. The variable selection in the GA model identified 19 important terms, 10 of them were positive and 9 were negative. The 10 GA positive terms coincide with the top 12 positives in the PLS (LOO) model, while the 9 GA negatives are the same as the top 9 PLS negatives.

PLS models generate loading plots to show which variables have the greatest influence on classification. The more distant from the origin are the variables, the greater effect they have on the division of classes. The loading plot for the first two components of the PLS (LOO) model is shown in Fig. 3. It is evident that terms $A_{11}(2)$, $A_{11}(4)$ and $A_{22}(2)$ are the most important for discriminating antigens ($\$DA_1$), while $C_{21}(1)$ overlaps with $\$DA_2$ (non-antigen class).

4. Discussion

We illustrate here a new approach to antigen identification based on the physicochemical properties of amino acids sequences. The method is alignment-independent and applies ACC preprocessing. The predictive ability of our

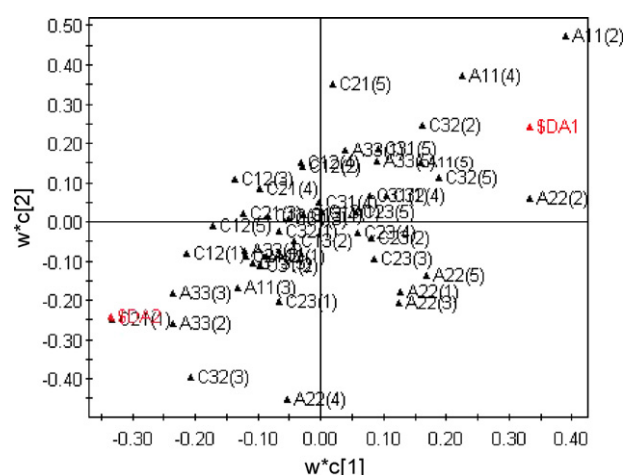


Fig. 3. Loading plot for the first two components of the PLS (LOO) model. The antigen and non-antigen classes are marked as $\$DA_1$ and $\$DA_2$, respectively.

models was tested by internal cross-validation and external validation on a test set. The highest prediction $accuracy$ was 83% for the cross-validation and 80% for the external test set. Many antigens in the test set were from bacterial species not included in the training set. The good predictive ability of the method suggests that antigens possess common, underlying physicochemical features which are independent of species and conventional global sequence similarity.

These models allow one to extract a physicochemical pattern common to antigen proteins. The most important terms for discrimination between antigens and non-antigens are $A_{11}(2)$, $A_{11}(4)$, $A_{22}(2)$ and $C_{21}(1)$. The first three are positive in the PLS (LOO) and GA models, while the last is negative. $A_{11}(2)$ and $A_{11}(4)$ are related to the auto covariance of the descriptor z_1 , which accounts for amino acid hydrophobicity. Hydrophilic amino acids (Asp, Asn, Glu, Arg, Lys, His, Gly, Gln, Ser, Thr and Cys) have positive z_1 values, while hydrophobic amino acids (Phe, Trp, Ile, Leu, Val, Met, Tyr and Pro) have negative z_1 values. Ala is amphiphilic with z_1 almost 0. As $A_{11}(2)$ and $A_{11}(4)$ are positive, in order to contribute to the antigen discrimination they should be positive, i.e. a product of two positives or two negatives. This means when in close proximity ($i+2$ or $i+4$) the amino acids should have similar hydrophobic properties. A lag of 4 is a well-known pattern for α -helices, where each amino acid corresponds to a 100° turn i.e., helices have 3.6 residues per turn. The side chains of amino acids at every first and fifth position are oriented in the same direction and make close contacts. They could be part of a conformational epitope.

Lag of 2 orients the side chains of the amino acids in opposite directions and it is associated with the β strand, a commonly occurring regular secondary structure in proteins. The linking loop between two parallel strands almost always has a right-handed crossover chirality, which is strongly favored by the inherent twist of the sheet. This linking loop often

contains a helical region, in which case it is called a β - α - β motif [9]. The equal importance of lags 2 and 4 for the antigen class may point to an overabundance of such motifs in antigen structures.

$A_{22}(2)$ is the auto covariance of the descriptor z_2 with lag of 2. The descriptor z_2 accounts for the size of the amino acids. Amino acids with a small size (Ala, Ser, Val, etc.) have negative z_2 values, amino acids with bulky side chains take positive ones. Every first and third amino acids in the antigen sequence should be of similar size. A pattern with amino acids of similar size and hydrophathy at i and $i + 2$ may be associated with the formation of surface areas with common antigenic properties. It is known that antigens bind to relatively flat combining sites on the complementarity-determining regions (CDRs) of antibodies [59,60]. An area of amino acids with similar size and hydrophathy is able to form a flat antigenic surface region complementary to the combining site of an antibody, thereby reducing the entropic penalty of the interaction.

The term $C_{21}(1)$ accounts for the cross covariance of z_2 and z_1 of successive amino acids. In the models derived here $C_{21}(1)$ has a negative coefficient. This implies that a combination of small-sized ($-z_2$) and hydrophilic ($+z_1$) or bulky ($+z_2$) and hydrophobic ($-z_1$) neighbors will contribute positively to antigen discrimination. Such a pattern is common to both B cell and T cell epitopes. Previous observations [61] suggest that Tyr and Trp frequently alternate with small amino acids such as Gly, Ala and Ser. This pattern of residues might allow maximum mobility of Tyr and Trp during complex formation. A similar pattern is observed in antigen-antibody complex HyHEL-5 [62] where Met32 is located between Tyr31 and Tyr33 and permits the side-chains of the flanking aromatics

to be exposed and potentially mobile. Recently, we found a similar pattern in the middle part of T-cell epitopes [63].

In this study, an alignment-independent method for antigen recognition has been applied to a set of known bacterial antigens. As this method is universal, it can also be applied to any other type of antigen: viral, tumor, protozoic, or fungal. There is a similarity between our objective and the screening for pharmacologically active compounds in drug discovery. The efficiency of antigen identification can be quantified in terms of enrichment factors compared to random screening. We seek to enrich significantly our selection with candidate antigens compared to the background frequency within the genome. It may also be possible to combine synergistically other *in silico* tools, such as prediction of gene expression and sub cellular location, to increase this enrichment even further. Methods that accurately predict candidate antigens will prove to be vital tools for the vaccinologist of tomorrow.

Acknowledgements

We thank Martin Blythe for helpful comments on the manuscript. The present study was supported in part by grants from the Royal Society, UK, and the Ministry of Education and Science, Bulgaria.

Appendix A

Training and test sets of bacterial antigens and non-antigens used in the study.

Species	Training set			
	Protein	Protection	Swiss-prot	Reference
<i>Bordetella pertussis</i>	Pertussis toxin S1 subunit	Yes	Q93V22	[14]
<i>Bordetella pertussis</i>	Pertactin	Yes	Q9S6N1	[14]
<i>Borrelia burgdorferi</i>	Outer membrane porin protein Oms28	Yes	P70854	[14]
<i>Borrelia burgdorferi</i>	Decorin-binding protein A	Yes	O50917	[14]
<i>Borrelia burgdorferi</i>	Outer surface protein A (ospA)	Yes	P14013	[15]
<i>Borrelia burgdorferi</i>	Outer surface protein B (ospB)	Yes	P17739	[15]
<i>Borrelia burgdorferi</i>	Outer surface protein C (ospC)	Yes	Q07337	[15]
<i>Brucella abortus</i>	Superoxide dismutase [Cu-Zn]	Yes	P15453	[14]
<i>Brucella abortus</i>	50S ribosomal protein L7/L12	Yes	P0A470	[14]
<i>Brucella melitensis</i>	25 kDa outer-membrane immunogenic protein	Yes	Q45321	[14]
<i>Campylobacter coli</i>	Flagellin FlaA	Yes	P27053	[16]
<i>Chlamydia trachomatis</i>	MOMP	Yes	Q46412	[17]
<i>Clostridium perfringens</i>	Phospholipase C	Yes	Q9RF12	[14]
<i>Clostridium perfringens</i>	Epsilon toxin	Yes	Q9RM68	[14]
<i>Clostridium tetani</i>	Tetanus toxin	Yes	Q9LA13	[14]
<i>Coccidioides immitis</i>	Urease	Yes	Q400Z7	[18]
<i>Coccidioides immitis</i>	Heat shock protein 60	Yes	O94110	[18]
<i>Coccidioides posadasii</i>	Aspartyl protease	Yes	Q3S565	[19]
<i>Corynebacterium pseudotuberculosis</i>	Phospholipase D	Yes	P20626	[14]
<i>Echinococcus granulosus</i>	EG95 host-protective vaccine antigen	Yes	Q24797	[20]
<i>Escherichia coli</i>	Heat-labile enterotoxin B subunit	Yes	Q93V32	[14]
<i>Escherichia coli</i>	Protein fimH	Yes	P08191	[14]

Appendix A (Continued)

Species	Training set			
	Protein	Protection	Swiss-prot	Reference
<i>Haemophilus influenzae</i>	Outer membrane protein P1	Yes	P43838	[21]
<i>Haemophilus influenzae</i>	Outer membrane protein P5	Yes	P45996	[14]
<i>Haemophilus influenzae</i>	Outer membrane protein P6	Yes	P10324	[14]
<i>Helicobacter pylori</i>	Citrate synthase	Yes	Q9ZN37	[22]
<i>Helicobacter pylori</i>	Urease B	Yes	Q7X3W5	[23]
<i>Helicobacter pylori</i>	Catalase	Yes	Q9ZKX5	[24]
<i>Helicobacter pylori</i>	NapA	Yes	Q9ZMJ1	[25]
<i>Helicobacter pylori</i>	10 kDa chaperonin Cytotoxicity-associated immunodominant antigen	Yes	P0A0R3	[14]
<i>Legionella pneumophila</i>	OmpS	Yes	Q9Z374	[26]
<i>Legionella pneumophila</i>	Heat shock protein 60	Yes	Q5ZXP3	[26]
<i>Legionella pneumophila</i>	Major secretory protein	Yes	P21347	[26]
<i>Listeria monocytogenes</i>	Listeriolysin O	Yes	Q9L5B9	[14]
<i>Listeria monocytogenes</i>	Protein p60	Yes	P21171	[14]
<i>Mycobacterium avium</i>	65 kDa protein	Yes	Q48900	[27]
<i>Mycobacterium avium</i>	Antigen 85B	Yes	Q06947	[27]
<i>Mycobacterium bovis</i>	MPB-83	Yes	P0A671	[28]
<i>Mycobacterium bovis</i>	Antigen 85A	Yes	P0A4V3	[27]
<i>Mycobacterium tuberculosis</i>	MPT64	Yes	P0A5Q4	[29]
<i>Mycobacterium tuberculosis</i>	PPE68	Yes	Q79F92	[30]
<i>Mycobacterium tuberculosis</i>	Ag85B	Yes	P13952	[31]
<i>Mycobacterium tuberculosis</i>	ESAT-6	Yes	P0A564	[32]
<i>Mycobacterium tuberculosis</i>	KatG	Yes	Q4TUW2	[32]
<i>Mycobacterium tuberculosis</i>	HBHA	Yes	P0A5P6	[32]
<i>Mycobacterium tuberculosis</i>	MPT-63	Yes	P0A5Q2	[33]
<i>Mycobacterium tuberculosis</i>	MPT-83	Yes	P0A670	[33]
<i>Mycobacterium tuberculosis</i>	PstS-3	Yes	P0A5Y2	[34]
<i>Mycobacterium tuberculosis</i>	PstS-2	Yes	O05870	[34]
<i>Mycobacterium tuberculosis</i>	Phosphate-binding protein 3	Yes	P0A5Y2	[14]
<i>Neisseria meningitidis</i>	NspA	Yes	P96943	[35]
<i>Neisseria meningitidis</i>	Transferrin binding protein TbpA	Yes	Q53348	[36]
<i>Neisseria meningitidis</i>	Transferrin binding protein TbpB	Yes	Q53990	[36]
<i>Pseudomonas aeruginosa</i>	Exotoxin A	Yes	P11439	[37]
<i>Pseudomonas aeruginosa</i>	Porin	Yes	P32722	[38]
<i>Pseudomonas aeruginosa</i>	PcrV	Yes	O30527	[39]
<i>Pseudomonas aeruginosa</i>	Outer membrane porin F	Yes	P13794	[14]
<i>Rickettsia tsutsugamushi</i>	56-kDa protein	Yes	Q6EZA6	[40]
<i>Shigella dysenteriae</i>	Shiga toxin B-chain	Yes	P69178	[14]
<i>Staphylococcus aureus</i>	Enterotoxin type A	Yes	P0A0L2	[14]
<i>Staphylococcus aureus</i>	Penicillin-binding protein 2'	Yes	Q7DHH4	[41]
<i>Staphylococcus aureus</i>	Clumping factor A	Yes	Q53653	[14]
<i>Streptococcus agalactiae</i>	Group B streptococcal sip protein	Yes	Q3K3Z5	[42]
<i>Streptococcus pneumoniae</i>	PspA	Yes	O34097	[43]
<i>Streptococcus pneumoniae</i>	PsaA	Yes	Q8VQ82	[43]
<i>Streptococcus pneumoniae</i>	Pneumolysin	Yes	P11990	[44]
<i>Streptococcus pneumoniae</i>	Choline binding protein A	Yes	Q8DN05	[44]
<i>Streptococcus pneumoniae</i>	PhpA	Yes	Q9AG74	[14]
<i>Streptococcus pyogenes</i>	Fibronectin-binding protein	Yes	Q01924	[14]
<i>Treponema pallidum</i>	Glycerophosphodiester phosphodiesterase	Yes	O30405	[45]
<i>Treponema pallidum</i>	tpr K	Yes	O83867	[46]
<i>Treponema pallidum</i>	TmpB	Yes	P19649	[47]
<i>Treponema pallidum</i>	Antigen TpF1	Yes	P16665	[14]
<i>Yersinia pestis</i>	V antigen	Yes	P21206	[48]
<i>Yersinia pestis</i>	CafI	Yes	P26948	[49]
<i>Bordetella pertussis</i>	Acetyl-coenzyme A carboxylase carboxyl transferase subunit alpha	No	Q7VX94	
<i>Bordetella pertussis</i>	Holo-[acyl-carrier-protein] synthase	No	Q7VWV9	
<i>Borrelia burgdorferi</i>	Outer surface protein D (ospD)	No	Q05051	
<i>Borrelia burgdorferi</i>	6-phosphogluconolactonase	No	O51240	
<i>Borrelia burgdorferi</i>	Acetate kinase	No	O51567	
<i>Borrelia burgdorferi</i>	Holo-[acyl-carrier-protein] synthase	No	O51043	

Appendix A (Continued)

Species	Training set			
	Protein	Protection	Swiss-prot	Reference
<i>Borrelia burgdorferi</i>	Adenine deaminase	No	Q50821	
<i>Brucella abortus</i>	Regulatory protein BvrR	No	Q67996	
<i>Brucella abortus</i>	Dihydrolipoamide succinyl transferase	No	Q85598	
<i>Brucella melitensis</i>	Translation initiation factor IF-2	No	Q8YEB3	
<i>Campylobacter coli</i>	Dihydrodipicolinate reductase	No	Q5HWX1	
<i>Chlamydia trachomatis</i>	Chorismate synthase	No	Q84373	
<i>Clostridium perfringens</i>	Shikimate dehydrogenase	No	Q8XMI8	
<i>Clostridium perfringens</i>	Chloramphenicol acetyltransferase	No	P26826	
<i>Clostridium tetani</i>	Transporter	No	Q890Y8	
<i>Coccidioides immitis</i>	Isocitrate lyase	No	Q96TP5	
<i>Coccidioides immitis</i>	Glyceraldehyde-3-phosphate dehydrogenase	No	Q8J1H3	
<i>Coccidioides posadasii</i>	Orotidine 5'-phosphate decarboxylase	No	Q4VWW3	
<i>Corynebacterium pseudotuberculosis</i>	3-dehydroquinase synthase	No	P96749	
<i>Echinococcus granulosus</i>	Paramyosin	No	P35417	
<i>Escherichia coli</i>	Pantoate-beta-alanine ligase	No	Q8X930	
<i>Escherichia coli</i>	UPF0053 inner membrane protein yoaE	No	P0AEC1	
<i>Haemophilus influenzae</i>	Biotin carboxylase	No	P43873	
<i>Haemophilus influenzae</i>	Nucleoside diphosphate kinase	No	P43802	
<i>Haemophilus influenzae</i>	UvrABC system protein A	No	Q4QNT9	
<i>Helicobacter pylori</i>	Chemotaxis protein cheY homolog	No	Q9ZM64	
<i>Helicobacter pylori</i>	Protein-export membrane protein secF	No	Q9ZJ65	
<i>Helicobacter pylori</i>	Carbonic anhydrase	No	Q25798	
<i>Helicobacter pylori</i>	Methylase	No	Q3S3S0	
<i>Helicobacter pylori</i>	Inorganic pyrophosphatase	No	Q8GK72	
<i>Legionella pneumophila</i>	Histidinol dehydrogenase	No	Q5X5W9	
<i>Legionella pneumophila</i>	Aconitate hydratase	No	Q5X4L7	
<i>Legionella pneumophila</i>	Triosephosphate isomerase	No	Q5ZRT6	
<i>Listeria monocytogenes</i>	Chemotaxis protein cheY	No	P0A4H5	
<i>Listeria monocytogenes</i>	Manganese transport system membrane protein mntC	No	Q8Y652	
<i>Mycobacterium avium</i>	Alanine and proline-rich secreted protein apa	No	Q48919	
<i>Mycobacterium avium</i>	Transposase	No	Q48909	
<i>Mycobacterium bovis</i>	Ribonuclease HII	No	Q7TXM7	
<i>Mycobacterium bovis</i>	Histidyl-tRNA synthetase	No	P67484	
<i>Mycobacterium tuberculosis</i>	PstS-1	No	P15712	
<i>Mycobacterium tuberculosis</i>	PE35	No	Q79F93	
<i>Mycobacterium tuberculosis</i>	Rv3878	No	Q69742	
<i>Mycobacterium tuberculosis</i>	Rv3879c	No	Q69743	
<i>Mycobacterium tuberculosis</i>	Tryptophan synthase beta chain	No	P66984	
<i>Mycobacterium tuberculosis</i>	Thioredoxin reductase	No	P52214	
<i>Mycobacterium tuberculosis</i>	Transcription elongation factor greA	No	P64279	
<i>Mycobacterium tuberculosis</i>	2,3-bisphosphoglycerate-dependent phosphoglycerate mutase	No	P0A5R6	
<i>Mycobacterium tuberculosis</i>	DNA gyrase subunit A	No	Q07702	
<i>Mycobacterium tuberculosis</i>	DNA gyrase subunit B	No	Q9L7L3	
<i>Mycobacterium tuberculosis</i>	Delta-aminolevulinic acid dehydratase	No	Q33357	
<i>Neisseria meningitidis</i>	Amylosucrase	No	Q84HD6	
<i>Neisseria meningitidis</i>	Chorismate synthase	No	Q9JY99	
<i>Neisseria meningitidis</i>	Argininosuccinate synthase	No	Q9JXC1	
<i>Pseudomonas aeruginosa</i>	Gentamicin 3'-acetyltransferase	No	P23181	
<i>Pseudomonas aeruginosa</i>	Biotin carboxylase	No	P37798	
<i>Pseudomonas aeruginosa</i>	Isocitrate lyase	No	Q9I0K4	
<i>Pseudomonas aeruginosa</i>	Glycosyl transferase alg8	No	Q52463	
<i>Rickettsia tsutsugamushi</i>	Rickettsia tsutsugamushi	No	Q53247	
<i>Shigella dysenteriae</i>	RepA	No	Q326M5	
<i>Staphylococcus aureus</i>	6-phosphogluconate dehydrogenase	No	Q931R3	
<i>Staphylococcus aureus</i>	Acetate kinase	No	Q931P6	
<i>Staphylococcus aureus</i>	Fructose-bisphosphate aldolase	No	Q6G7I5	
<i>Streptococcus agalactiae</i>	Arginine deiminase	No	Q8E2K0	
<i>Streptococcus pneumoniae</i>	Dihydrodipicolinate synthase	No	Q97R25	
<i>Streptococcus pneumoniae</i>	Tyrosine recombinase xerC	No	Q7ZAK7	
<i>Streptococcus pneumoniae</i>	Initiation-control protein yabA	No	Q97R89	

Appendix A (Continued)

Species	Training set			
	Protein	Protection	Swiss-prot	Reference
<i>Streptococcus pneumoniae</i>	Capsular polysaccharide synthesis protein	No	O07341	
<i>Streptococcus pneumoniae</i>	Transposase	No	O33754	
<i>Streptococcus pyogenes</i>	Alanine racemase	No	Q99Y98	
<i>Treponema pallidum</i>	TmpA	No	P07643	
<i>Treponema pallidum</i>	TmpC	No	P29724	
<i>Treponema pallidum</i>	Methionine aminopeptidase	No	O83814	
<i>Treponema pallidum</i>	Chemotaxis protein cheA	No	P96123	
<i>Yersinia pestis</i>	Uronate isomerase	No	Q8ZIC6	
<i>Yersinia pestis</i>	Flavoprotein wrbA	No	Q8ZF61	
Species	Test set			
	Protein	Protection	Swiss-prot	Reference
<i>Escherichia coli</i>	FepA	Yes	P05825	[50]
<i>Escherichia coli</i>	Cell division inhibitor	Yes	P0AFZ6	[50]
<i>Escherichia coli</i>	Colicin I receptor precursor	Yes	P17315	[50]
<i>Klebsiella pneumoniae</i>	OmpA	Yes	P24017	[50]
<i>Klebsiella pneumoniae</i>	OmpK17	Yes	Q48427	[50]
<i>Klebsiella pneumoniae</i>	OmpK36	Yes	Q48473	[50]
<i>Vibrio parahaemolyticus</i>	Hypothetical protein VP2716	Yes	Q87L97	[50]
<i>Salmonella typhimurium</i>	OmpW	Yes	Q8ZP50	[50]
<i>Mycobacterium tuberculosis</i>	MPT51/MPB51 antigen	Yes	P0A4V6	[51]
<i>Mycobacterium tuberculosis</i>	CFP6	Yes	P0A5P2	[51]
<i>Mycobacterium tuberculosis</i>	CFP10	Yes	P0A566	[51]
<i>Mycobacterium tuberculosis</i>	Mtb10.4	Yes	P0A568	[51]
<i>Mycobacterium tuberculosis</i>	Mtb8.4	Yes	O50430	[51]
<i>Mycobacterium tuberculosis</i>	Mtb12	Yes	P0A5P8	[51]
<i>Mycobacterium tuberculosis</i>	Mtb9.9	Yes	P0A570	[51]
<i>Mycobacterium tuberculosis</i>	Mtb32A	Yes	O07175	[51]
<i>Mycobacterium tuberculosis</i>	PPE family protein (Mtb39)	Yes	Q7D8M9	[51]
<i>Mycobacterium tuberculosis</i>	PPE family protein (Mtb41)	Yes	Q79FV1	[51]
<i>Mycobacterium tuberculosis</i>	14 kDa antigen (hspX)	Yes	P0A5B7	[51]
<i>Mycobacterium tuberculosis</i>	Phosphate-binding protein 1 (PstS-1)	Yes	P15712	[51]
<i>Mycobacterium tuberculosis</i>	Putative lipoprotein lppX (LppX)	Yes	P65306	[51]
<i>Streptococcus agalactiae</i>	Hypothetical protein	Yes	Q9ZHG7	[52]
<i>Streptococcus pneumoniae</i>	Putative endo-beta-N-acetylglucosaminidase [Precursor]	Yes	P59206	[52]
<i>Streptococcus pneumoniae</i>	1,4-beta-N-acetylmuramidase [Precursor]	Yes	Q9Z4J8	[52]
<i>Bacillus anthracis</i>	PXO2-08	Yes	Q9RN24	[52]
<i>Klebsiella pneumoniae</i>	Aminoglycoside 3'-phosphotransferase	No	P00552	
<i>Klebsiella pneumoniae</i>	Pullulanase secretion protein pulS [Precursor]	No	P20440	
<i>Klebsiella pneumoniae</i>	His operon leader peptide	No	Q48439	
<i>Mycobacterium tuberculosis</i>	6-phosphogluconolactonase	No	P63338	
<i>Mycobacterium tuberculosis</i>	Aminoglycoside 2'-N-acetyltransferase	No	P0A5N0	
<i>Mycobacterium tuberculosis</i>	Acyl-CoA dehydrogenase fadE12	No	P71539	
<i>Mycobacterium tuberculosis</i>	Isocitrate lyase	No	P0A5H3	
<i>Mycobacterium tuberculosis</i>	Acetate kinase	No	P63409	
<i>Mycobacterium tuberculosis</i>	Meromycolate extension acyl carrier protein	No	P0A4W6	
<i>Mycobacterium tuberculosis</i>	Adenosine deaminase	No	P63907	
<i>Mycobacterium tuberculosis</i>	L-asparagine permease 1	No	O33261	
<i>Mycobacterium tuberculosis</i>	Potassium-transporting ATPase A chain	No	P65209	
<i>Mycobacterium tuberculosis</i>	Biotin synthase	No	P0A506	
<i>Mycobacterium tuberculosis</i>	Fumarate reductase flavoprotein subunit	No	P64174	
<i>Mycobacterium tuberculosis</i>	DNA translocase ftsK	No	O33290	
<i>Mycobacterium tuberculosis</i>	Galactokinase	No	P96910	
<i>Mycobacterium tuberculosis</i>	Hemoglobin-like protein HbN	No	P0A592	
<i>Mycobacterium tuberculosis</i>	Phosphoglucosamine mutase	No	O06258	
<i>Mycobacterium tuberculosis</i>	Nitrogen regulatory protein P-II	No	P64249	
<i>Mycobacterium tuberculosis</i>	Serine hydroxymethyltransferase 2	No	O53615	
<i>Mycobacterium tuberculosis</i>	Phosphoheptose isomerase	No	P0A604	

Appendix A (Continued)

Species	Test set			
	Protein	Protection	Swiss-prot	Reference
<i>Mycobacterium tuberculosis</i>	Porphobilinogen deaminase	No	P64336	
<i>Mycobacterium tuberculosis</i>	Translation initiation factor IF-1	No	P0A5H5	
<i>Escherichia coli</i>	Capsule polysaccharide export protein kpsC	No	P42217	
<i>Escherichia coli</i>	Bicyclomycin resistance protein	No	P28246	
<i>Escherichia coli</i>	Isoaspartyl dipeptidase	No	P39377	
<i>Escherichia coli</i>	FMN reductase	No	P80644	
<i>Escherichia coli</i>	Leu operon leader peptide	No	P0AD79	
<i>Salmonella typhimurium</i>	Universal stress protein A	No	Q8ZLD7	
<i>Salmonella typhimurium</i>	Phosphoglycolate phosphatase	No	Q8ZLK5	
<i>Salmonella typhimurium</i>	Formate-dependent nitrite reductase	No	Q8ZKF4	
<i>Vibrio parahaemolyticus</i>	Nucleoside diphosphate kinase	No	Q87S20	
<i>Vibrio parahaemolyticus</i>	Acetylornithine deacetylase	No	P59601	
<i>Vibrio parahaemolyticus</i>	Chorismate synthase	No	Q87MM9	
<i>Streptococcus agalactiae</i>	Mannonate dehydratase	No	Q3K203	
<i>Streptococcus agalactiae</i>	Transporter, BCCT family protein	No	Q3D265	
<i>Streptococcus agalactiae</i>	Homoserine kinase	No	Q3D272	
<i>Streptococcus pneumoniae</i>	Dihydropteroate synthase	No	P05382	
<i>Streptococcus pneumoniae</i>	Isoleucyl-tRNA synthetase	No	Q9ZHB3	
<i>Streptococcus pneumoniae</i>	Pneumolysin	No	Q2XU25	
<i>Streptococcus pneumoniae</i>	Methionyl-tRNA formyltransferase	No	Q97PA6	
<i>Streptococcus pneumoniae</i>	Formamidopyrimidine-DNA glycosylase	No	Q97R61	
<i>Bacillus anthracis</i>	Iron compound ABC transporter, permease protein	No	Q6HR47	
<i>Bacillus anthracis</i>	Major facilitator family transporter	No	Q6HR65	
<i>Bacillus anthracis</i>	Histidinol dehydrogenase	No	Q81T62	

References

- [1] Rappuoli R. Reverse vaccinology, a genome-based approach to vaccine development. *Vaccine* 2001;19:2688.
- [2] Pizza M, Scarlato V, Masignani V, Giuliani MM, Arico B, Comanducci M, et al. Identification of vaccine candidates against serogroup B meningococcus by whole-genome sequencing. *Science* 2000;287:1816.
- [3] Mora M, Veggi D, Santini L, Pizza M, Rappuoli R. Reverse vaccinology. *Drug Discov Today* 2003;8:459.
- [4] Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol* 1990;215:403.
- [5] Lipman DJ, Pearson WR. Rapid and sensitive protein similarity searches. *Science* 1985;227:1435.
- [6] Nakai K, Kanehisa M. Expert system for predicting protein localization sites in gram-negative bacteria. *Proteins* 1991;11:95.
- [7] Bendtsen JD, Nielsen H, von Heijne G, Brunak S. Improved prediction of signal peptides: SignalP 3.0. *J Mol Biol* 2004;340:783.
- [8] Accelrys GCG 11.0, July 2005.
- [9] Petsko GA, Ringe D. Protein structure and function. Blackwell Publishing; 2004.
- [10] Wold S, Jonsson J, Sjöström M, Sandberg M, Rännar S. DNA and peptide sequences and chemical processes multivariately modeled by principal component analysis and partial least-squares projections to latent structures. *Anal Chim Acta* 1993;277:239.
- [11] Andersson PM, Sjöström M, Lundstedt T. Preprocessing peptide sequences for multivariate sequence-property analysis. *Chemometr Intell Lab* 1998;42:41.
- [12] Nyström Å, Andersson PM, Lundstedt T. Multivariate data analysis of topographically modified α -melanotropin analogues using auto and cross auto covariances (ACC). *Quant Struct-Act Rel* 2000;19:264.
- [13] Lapinsh M, Gutcaits A, Prusis P, Post C, Lundstedt T, Wikberg JES. Classification of G-protein coupled receptors by alignment-independent extraction of principal chemical properties of primary amino acid sequences. *Protein Sci* 2002;11:795.
- [14] Mayers C, Duffield M, Rowe S, Miller J, Lingard B, Hayward S, et al. Analysis of known bacterial protein vaccine antigens reveals biased physical properties and amino acid composition. *Comp Funct Genom* 2003;4:468.
- [15] Probert WS, LeFebvre RB. Protection of C3H/HeN mice from challenge with *Borrelia burgdorferi* through active immunization with OspA, OspB, or OspC, but not with OspD or the 83-kilodalton antigen. *Infect Immun* 1994;62:1920.
- [16] Lee LH, Burg III E, Baqar S, Bourgeois AL, Burr DH, Ewing CP, et al. Evaluation of a truncated recombinant flagellin subunit vaccine against *Campylobacter jejuni*. *Infect Immun* 1999;67:5799.
- [17] Igietsme JU, Murdin A. Induction of protective immunity against *Chlamydia trachomatis* genital infection by a vaccine based on major outer membrane protein-lipophilic immune response-stimulating complexes. *Infect Immun* 2000;68:6798.
- [18] Li K, Yu J, Hung C, Lehmann PF, Cole GT. Recombinant urease and urease DNA of *Coccidioides immitis* elicit an immunoprotective response against coccidioidomycosis in mice. *Infect Immun* 2001;69:2878.
- [19] Tarcha EJ, Basrur V, Hung C, Gardner MJ, Cole GT. A recombinant aspartyl protease of *Coccidioides posadasii* induces protection against pulmonary coccidioidomycosis in mice. *Infect Immun* 2006;74:516.
- [20] Woollard DJ, Gauci CG, Heath DD, Lightowlers MW. Protection against hydatid disease induced with the EG95 vaccine is associated with conformational epitopes. *Vaccine* 2001;19:498.
- [21] Bolduc GR, Bouchet V, Jiang R, Geisselsoder J, Truong-Bolduc QC, Rice PA, et al. 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* 2000;68:4505.

- [22] Dunkley ML, Harris SJ, McCoy RJ, Musicka MJ, Evers FM, Beagley LG, et al. Protection against *Helicobacter pylori* infection by intestinal immunization with a 50/52-kDa subunit protein. *FEMS Immunol Med Mic* 1999;24:221.
- [23] Ermak TH, Giannasca PJ, Nichols R, Myers GA, Nedrud J, Weltzin R, et al. Immunization of mice with urease vaccine affords protection against *Helicobacter pylori* infection in the absence of antibodies and is mediated by MHC class II-restricted responses. *J Exp Med* 1998;188:2277.
- [24] Radcliff FJ, Hazell SL, Kolesnikow T, Doidge C, Lee A. Catalase, a novel antigen for *Helicobacter pylori* vaccination. *Infect Immun* 1997;65:4668.
- [25] Satin B, Del Giudice G, Della Bianca V, Dusi S, Laudanna C, Tonello F, et al. The neutrophil-activating protein (HP-NAP) of *Helicobacter pylori* is a protective antigen and a major virulence factor. *J Exp Med* 2000;191:1467.
- [26] Weeratna R, Stamler DA, Edelstein PH, Ripley M, Marrie T, Hoskin D, et al. Human and guinea pig immune responses to *Legionella pneumophila* protein antigens OmpS and Hsp60. *Infect Immun* 1994;62:3454.
- [27] Velaz-Faircloth M, Cobb AJ, Horstman AL, Henry SC, Frothingham R. Protection against *Mycobacterium avium* by DNA vaccines expressing mycobacterial antigens as fusion proteins with green fluorescent protein. *Infect Immun* 1999;67:4243.
- [28] Chambers MA, Vordermeier HM, Whelan A, Commander N, Tascon R, Lowrie D, et al. Vaccination of mice and cattle with plasmid DNA encoding the *Mycobacterium bovis* antigen MPB83. *Clin Infect Dis* 2000;30(Suppl. 3):S283.
- [29] Delogu G, Howard A, Collins FM, Morris SL. DNA vaccination against tuberculosis: Expression of a ubiquitin-conjugated tuberculosis protein enhances antimycobacterial immunity. *Infect Immun* 2000;68:3097.
- [30] Demangel C, Brodin P, Cockle PJ, Brosch R, Majlessi L, Leclerc C, et al. Cell envelope protein PPE68 contributes to Mycobacterium tuberculosis RD1 immunogenicity independently of a 10-kilodalton culture filtrate protein and ESAT-6. *Infect Immun* 2004;72:2170.
- [31] Kamath AT, Groat NL, Bean AGD, Britton WJ. Protective effect of DNA immunization against mycobacterial infection is associated with the early emergence of interferon-gamma (IFN- γ)-secreting lymphocytes. *Clin Exp Immunol* 2000;120:476.
- [32] Li Z, Howard A, Kelley C, Delogu G, Collins F, Morris S. Immunogenicity of DNA vaccines expressing tuberculosis proteins fused to tissue plasminogen activator signal sequences. *Infect Immun* 1999;67:4780.
- [33] Morris S, Kelley C, Howard A, Li Z, Collins F. The immunogenicity of single and combination DNA vaccines against tuberculosis. *Vaccine* 2000;18:2155.
- [34] Tanghe A, Lefevre P, Denis O, D'Souza S, Braibant m, Lozes E, et al. Immunogenicity and protective efficacy of tuberculosis DNA vaccines encoding putative phosphate transport receptors. *J Immunol* 1999;162:1113.
- [35] Martin D, Cadieux N, Hamel J, Brodeur BR. Highly conserved *Neisseria meningitidis* surface protein confers protection against experimental infection. *J Exp Med* 1997;185:1173.
- [36] West D, Reddin K, Matheson M, Heath R, Funnell S, Hudson M, et al. Recombinant *Neisseria meningitidis* transferrin binding protein A protects against experimental meningococcal infection. *Infect Immun* 2001;69:1561.
- [37] Denis-Mize KS, Price BM, Baker NR, Galloway DR. Analysis of immunization with DNA encoding *Pseudomonas aeruginosa* exotoxin A. *FEMS Immunol Med Mic* 2000;27:147.
- [38] Gilleland Jr HE, Gilleland LB, Metthews-Greer JM. Outer membrane protein F preparation of *Pseudomonas aeruginosa* as a vaccine against chronic pulmonary infection with heterologous immunotype strains in a rat model. *Infect Immun* 1988;56:1017.
- [39] Holder IA, Neely AN, Frank DW. PcrV immunization enhances survival of burned *Pseudomonas aeruginosa*-infected mice. *Infect Immun* 2001;69:5908.
- [40] Seong S, Huh MS, Jang WJ, Park SG, Kim JG, Woo SG, et al. Induction of homologous immune response to *Rickettsia tsutsugamushi* Boryong with a partial 56-kilodalton recombinant antigen fused with the maltose-binding protein MBP-Bor56. *Infect Immun* 1997;65:1541.
- [41] Ohwada A, Sekiya M, Hanaki H, Arai KK, Nagaoka I, Hori S, et al. DNA vaccination by mecA sequence evokes an antibacterial immune response against methicillin-resistant *Staphylococcus aureus*. *J Antimicrob Chemother* 1999;44:767.
- [42] Brodeur BR, Boyer M, Charlebois I, Hamel J, Couture F, Rioux CR, et al. Identification of group B streptococcal sip protein which elicits cross-protective immunity. *Infect Immun* 2000;68:5610.
- [43] Briles DE, Ades E, Paton JC, Sampson JS, Carlone GM, Huebner RC, et al. Intranasal immunization of mice with a mixture of the pneumococcal proteins PsaA and PspA is highly protective against nasopharyngeal carriage of *Streptococcus pneumoniae*. *Infect Immun* 2000;68:796.
- [44] Ogunniji AD, Woodrow MC, Poolman JT, Paton JC. Protection against *Streptococcus pneumoniae* elicited by immunization with pneumolysin and CbpA. *Infect Immun* 2001;69:5997.
- [45] Cameron CE, Castro C, Lukehart SA, van Voorhis WC. Function and proteic capacity of *Treponema pallidum* subsp *pallidum* glycerophosphodiester phosphodiesterase. *Infect Immun* 1998;66:5763.
- [46] Centurion-Lara A, Castro C, Barrett L, Cameron C, Mostowfi M, van Voorhis WC, et al. *Treponema pallidum* major sheath protein homologue Tpr K is a target of opsonic antibody and the protective immune response. *J Exp Med* 1999;189:647.
- [47] Wicher K, Schouls LM, Wicher V, van Embden JDA, Nakeeb SS. Immunization of guinea pigs with recombinant TmpB antigen induces protection against challenge infection with *Treponema pallidum* Nichols. *Infect Immun* 1991;59:4343.
- [48] Anderson Jr GW, Leary SEC, Williamson ED, Titball RW, Welkos SL, Worsham PL, et al. Recombinant V antigen protects mice against pneumonic and bubonic plague caused by F1-capsule-positive and -negative strains of *Yersinia pestis*. *Infect Immun* 1996;64:4580.
- [49] Elvin SJ, Eyles JE, Howard KA, Ravichandran E, Somavarappu S, Alpar HO, et al. Protection against bubonic and pneumonic plague with a single dose microencapsulated sub-unit vaccine. *Vaccine* 2006;24:4433.
- [50] Kurupati P, The BK, Kumarasinghe G, Poh CL. Identification of vaccine candidate antigens of an ESBL producing *Klebsiella pneumoniae* clinical strain by immunoproteome analysis. *Proteomics* 2006;6:836.
- [51] Andersen P, Doherty TM. TB subunit vaccines—putting the pieces together. *Microbes Infect* 2005;7:911.
- [52] Wizemann TM, Heinrichs JH, Adamou JE, Erwin AL, Kunsch C, Choi GH, et al. Use of a whole genome approach to identify vaccine molecules affording protection against *Streptococcus pneumoniae* infection. *Infect Immun* 2001;69:1593.
- [53] Hellberg S, Sjöström M, Skagerberg B, Wold S. Peptide quantitative structure–activity relationships, a multivariate approach. *J Med Chem* 1987;30:1126.
- [54] Sandberg M, Eriksson L, Jonsson J, Sjöström M, Wold S. New chemical descriptors relevant for the design of biologically active peptides A multivariate characterization of 87 amino acids. *J Med Chem* 1998;41:2481.
- [55] Leardi R, Boggia R, Terrile M. Genetic algorithms as a strategy for feature-selection. *J Chemometrics* 1992;6:267.
- [56] MDL QSAR 2.2. 14600 Catalina St, San Leandro CA 94577.
- [57] SIMCA 8.0. Umetrics UK Ltd., Wokingham Road, RG42 1PL, Bracknell, UK.
- [58] Bradley AP. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recogn* 1997;30:1145.

- [59] MacCallum RM, Martin ACR, Thornton JM. Antibody-antigen interactions: contact analysis and binding site topography. *J Mol Biol* 1996;262:732.
- [60] Collis AVJ, Brouwer AP, Martin ACR. Analysis of the antigen combining site: correlations between length and sequence composition of the hypervariable loops and the nature of the antigen. *J Mol Biol* 2003;325:337.
- [61] Mian IS, Bradwell AR, Olson AJ. Structure, function and properties of antibody binding sites. *J Mol Biol* 1991;217:133.
- [62] Chothia C, Lesk AM, Tramontano A, Levitt M, Smith-Gill SJ, Air G, et al. Conformations of immunoglobulin hypervariable regions. *Nature* 1989;342:877.
- [63] Doytchinova IA, Flower DR. Modeling the peptide – T-cell receptor interaction by the comparative molecular similarity analysis – soft independent modeling of class analogy technique. *J Med Chem* 2006;49:2193.

Author's personal copy