Bioinformatic Approach for Identifying Parasite and Fungal Candidate Subunit Vaccines

Article *in* The Open Vaccine Journal · September 2008

DOI: 10.2174/1875035400801010022

Irini A Doytchinova
Medical University of Sofia
161 PUBLICATIONS 5,295 CITATIONS

SEE PROFILE

Darren Flower
Independent
288 PUBLICATIONS 13,903 CITATIONS

SEE PROFILE

Open Access

Bioinformatic Approach for Identifying Parasite and Fungal Candidate Subunit Vaccines

Irini A. Doytchinova*[,1] and Darren R. Flower[2]

[1]*Faculty of Pharmacy, Medical University of Sofia, 2 Dunav st., 1000 Sofia, Bulgaria*

[2]*The Jenner Institute, University of Oxford, Compton, Berkshire RG20 7NN, UK*

Abstract: *In silico* genome analysis enables systematic identification of potential antigens within a pathogen. Of primary importance is the accuracy of computer algorithms used for antigen prediction. Most bioinformatics tools are based on sequence alignment and are not able to predict truly novel antigenic proteins which lack similarity to existing antigens or which encode antigenicity in a cryptic manner. To surmount such obstacles, we have recently developed an alignment-free approach to *in silico* antigen identification, based on the auto cross covariance (ACC) transformation of protein sequences into uniform vectors of principal amino acid properties. Here, we apply this approach to finding parasite and fungal immunoprotective antigens. The models derived in this study demonstrate good predictive ability with 78% to 97% accuracy under internal cross validation in 7 groups. Under external validation, they gave 69% sensitivity ranking the true immunoprotective proteins in the first 25% of their proteomes.

Keywords: Protective antigens, parasite antigens, fungal antigens, discriminant analysis, partial least squares, cross validation, ROC statistics, accuracy, sensitivity, specificity.

INTRODUCTION

Diseases of helminth parasite and protozoan origin cause significant mortality and morbidity in tropical and subtropical countries. The World Health Organization estimate 3.5 billion people currently suffer parasitic infection [1]. Such infections are common in pastoral regions of Asia, Latin America, and Africa; they are less prevalent in industrialized countries. Parasites invade the body *via* the mouth or skin. Buccal parasites are swallowed and can remain in the intestine or egress from it through the intestinal wall, invading other organs. Other parasites bore directly through the skin or enter *via* bites from infected insects (the vector). Many parasites, particularly single-celled parasites, reproduce within the host. Parasites can have complex life cycles, producing eggs or larvae that reside in the environment or within an insect vector before becoming infective. Controlling disease-carrying invertebrate vectors (i.e. mosquitoes) remains problematic. No vaccines are currently available to prevent or control the spread of parasitic diseases.

Compared to parasitic disease, fungal infections are less harmful but more tenacious. Certain fungi (such as *Candida*) typically colonise outer body and mucosal surfaces, such as the intestines. Overtly harmless, such fungi can induce local infections of the skin, nails, mouth, vagina, or sinuses: they are seldom serious, except in those with impaired immunity. Occasionally, infections occur when habitual suppression of fungi is disturbed. For example, bacteria found in the vagina or alimentary canal usually limit fungal growth. Antibiotics can kill such helpful bacteria, allowing fungi to grow unchecked, and thus causing symptoms, which are usually mild. As the bacterial population is restored, so is the balance, and symptoms are resolved. Certain fungal infections - histoplasmosis, coccidioidomycosis, or blastomycosis - may be serious in otherwise healthy people. As fungal infections can develop slowly, months or years can pass before treatment is sought. Fungal infection can be very aggressive for those with impaired immunity: those on immunosuppressive anti-tumour chemotherapy, organ transplants, or who have AIDS. Disease spreads quickly to other organs and may cause death.

Vaccines are among the most efficacious prophylactic treatments for infectious disease. Vaccines can be divided into two groups: living and non-living. Living vaccines are typically attenuated pathogens which induce immunity by mimicking aspects of natural infection. Non-living vaccines are either whole killed pathogens or pathogen components (subunit vaccines) able to induce protective immunity [2]. Subunit vaccines have several advantages compared to whole pathogen vaccines. They are stable, have long shelf-lives, and do not revert to a pathogenic state [3]. Their preparation is, however, time-, source-, and labor-consuming, and some pathogens cannot be cultivated.

Genomics has opened up a new era in vaccine development. Burgeoning pathogenic genome information permits the development of vaccines *in silico*, without needing to cultivate the micro-organism. This approach, called reverse vaccinology [4], when used with advanced molecular biology technology, enables the systematic identification of potential antigens within a pathogen based on the *in silico* analysis of its genome. Of primary importance in this approach is the accuracy of the computer algorithms used for antigen identification. Many bioinformatics tools can identify surface-associated or outer membrane proteins, signal

*Address correspondence to this author at the Faculty of Pharmacy, Medical University of Sofia, 2 Dunav st., 1000 Sofia, Bulgaria,
E-mail: idoytchinova@pharmfac.net

1875-0354/08 2008 Bentham Open

proteins, lipoproteins, or host-cell binding domains [5-8]. Sequence alignment-based analysis is unable to predict truly novel antigenic proteins which lack similarity with existing antigens or which encode antigenicity in a subtle and cryptic manner. In order to address this, recently we have developed an alignment-free method for antigen prediction. This approach is based on the auto cross covariance (ACC) transformation of protein sequences into uniform vectors of principal amino acid properties. It has been applied to bacterial, viral and tumour antigens and the resulting models showed prediction accuracy of 70% to 89% [9,10]. The models were implemented in the VaxiJen server, and are thus freely available: http://www.jenner.ac.uk/VaxiJen.

In the present study, we apply this approach to sets of parasite and fungal antigens and non-antigens in order to develop models able to distinguish immunoprotective proteins from the general microbial proteome.

MATERIALS AND METHODOLOGY

Antigens and Non-Antigens Datasets

Antigens of parasite and fungal origin were identified from the literature. The parasite antigen subset consisted of 117 proteins, while the fungal antigen subset included 33 proteins. A protein was identified as an immunoprotective antigen if it, or part thereof, or its corresponding DNA, has been shown to be protective following immunization in an appropriate animal model. The non-antigen subsets were constructed to mirror the antigens subsets. For each antigen, a non-antigen protein was randomly selected from the same species. Proteomes and protein sequences were obtained from the UniProt Knowledgebase on the ExPASy Proteomics Server [11]. The proteins used in the study are given in Supplementary Materials 1 (parasite proteins) and 2 (fungal proteins).

z-Descriptors and ACC Transformations

The *z*-descriptors, as defined by Hellberg and collaborators [12], summarize the principal physicochemical properties of the amino acids. These descriptors were derived by principal component analysis of a data matrix consisting of 29 molecular descriptors. The first principal component (z_1) reflects the hydrophobicity of amino acids, the second (z_2) their size, and the third (z_3) their polarity. By arranging the *z* values according to the amino acid sequence, it is possible to quantify the structural variations numerically within a series of related proteins. In the present study the z_1, z_2 and z_3 descriptors were used to describe the protein sequences.

To make uniform the length of the proteins used in the present study, an auto-cross covariance (ACC) transformation was used [13]. Auto covariance *Ajj(l)* and cross covariance *Cjk(l)* were calculated according to Eqs. (1) and (2), respectively:

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

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

Index *j* refers to the *z*-descriptors (*j* = 1-3), *n* is the number of amino acids in a sequence, index *i* ponts the amino acid position (*i* = 1, 2, …, *n*) and *l* is the lag (*l* = 1, 2, …, *L*). Short lags (*L* = 5) have been chosen as only the influence of the close amino acid proximity was investigated. The subsets of antigens and non-antigens were transformed into matrices with 45 variables (3^2 x 5) each.

Discriminant Analysis by Partial Least Squares (DA – PLS)

Sets consisted of immunoprotective antigens and non-antigens were subjected to two-class discriminant analysis by partial least squares (DA-PLS) using SIMCA-P 8.0 [14]. The optimum number of components was selected by adding components until the next added component explained less than 10% of the variance. The predictive accuracy of the models was assessed by cross validation in 7 groups. The correctly predicted antigens and non-antigens were defined as true positives (TP) and true negatives (TN), respectively. The incorrectly predicted antigen and non-antigens were defined as false negatives (FN) and false positives (FP), respectively. *Sensitivity* [TP/(TP + FN)] and *1 – specificity* [FP/(TN + FP)] were calculated at different threshold and plotted to generate Receiver Operating Characteristic (*ROC*) curves [15]. The area under the curve (A_{ROC}) is a quantitative measure of the predictability of the models. It varies from 0.5 (random prediction) to 1.0 (perfect prediction). *Accuracy* [(TP + TN)/total] of the models at different thresholds also was calculated.

VaxiJen Server

The new models derived here were incorporated into a revised VaxiJen server. VaxiJen is implemented in Perl, with an interface written in HTML. It identifies bacterial, viral, tumour, parasite and fungal antigens. Protein sequences are uploaded as single or multiple files in plain or fasta format. The results page returns immunoprotective antigen probability (as a fraction of unity) for each protein and a statement of antigen status (“Probable Antigen” or “Probable Non-antigen”). Optional, ACC values are displayed on the results page.

RESULTS

Model for Parasite Antigen Prediction

The parasite training set consists of 117 known immunoprotective antigens and 117 non-antigens. Each amino acid in a protein sequence was described by a set of *z* descriptors: z_1 (hydrophobicity), z_2 (size) and z_3 (polarity). To fit proteins in a uniform length, ACC transformations were applied. A matrix of 45 columns and 234 rows was generated. Proteins were divided into two classes and discriminant analysis was applied using PLS method. The predictive ability of the derived model was tested by cross validation in 7 groups and assessed by A_{ROC}, highest *accuracy*, *sensitivity* and *specificity* (Table **1**). The parasite model has A_{ROC} value of 0.851 (Fig. **1**), highest *accuracy* 78% at threshold 0.5, the *sensitivity* and *specificity* at this threshold are 73% and 84%, respectively.

Model for Fungal Antigen Prediction

The fungal training set consists of 33 known immunoprotective antigens and 33 randomly selected non-antigens from the same species. The protein sequences were described by *z* descriptors and processed by ACC transformation. The generated matrix underwent discriminant analysis by PLS. The model was cross validated in 7 groups and showed A_{ROC} = 0.994 (Fig. **2**), highest *accuracy* 97% at threshold 0.5, *sensitivity* 97% and *specificity* 97% at this threshold (Table **1**).

Table 1. ROC Statistics of the Parasite and Fungal Models

Model	Test Set	A_{ROC}	Threshold %	Accuracy %	Sensitivity %	Specificity %
parasite	parasite	0.851	0.5	78	73	84
	fungal	0.550	0.5	53	71	36
fungal	fungal	0.994	0.5	97	97	97
	parasite	0.614	0.5	59	70	48

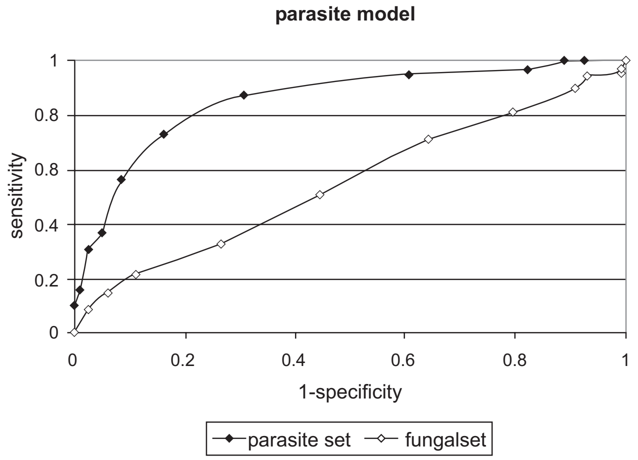

Fig. (1). ROC curve of the parasite model.

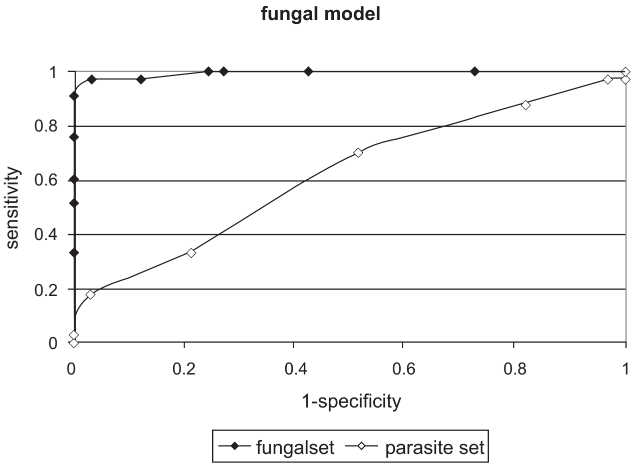

Fig. (2). ROC curve of the fungal model.

Cross Comparative Analysis

To examine the kingdom specificity of the derived models, a cross comparative analysis was performed. The fungal set was used as a test set for the parasite model and the parasite set was used as a test set for the fungal model. Predictions were compared in terms of their ROC statistics (Table **1**, Figs. **1** and **2**). It is clearly evident that the models are strictly kingdom specific. Practically, they lack predictive ability when applied to a test set with a different origin.

External Validation

To validate the predictive ability of the derived models, we collected a test set of known parasite and fungal antigens with immunoprotective properties published in 2007 and 2008: this set was not part of the training data. The test set comprised 25 parasite and 1 fungal immunoprotective antigens and is given in Table **2**.

Table 2. Test Set for External Validation

Species [Reference]	Immunogen	Predicted Score	Proteins in Swiss-Prot	Outliers	Rank %
Brugia malayi [16]	Q04009	0.608	184	40	12
Fasciola hepatica [17]	Q17TZ3	0.209	23	15	
Leishmania major [18,19]	Q4Q1I2	0.309	40	12	
	Q4Q1I3	0.316			0
	A9LJZ6	0.361			
	Q4Q271	0.881			
	Q9NKJ7	0.446			
Schistosoma bovis [20]	Q6IX09	0.653	14	3	18
	Q6IX10	0.184			
Schistosoma japonicum [21]	P43157	0.485	37	20	
	P42665	0.801			0
Plasmodium berghei [22]	Q6W9J3	0.229	16	4	
Plasmodium falciparum [23,24]	O96930	0.602	204	38	25
	O15863	0.739			12
	O15862	0.783			10
	O15856	0.790			10
	O15860	0.798			10
	O15859	0.812			10
	O15857	0.835			8
	O15854	0.862			8
	O15853	0.881			8
	O15861	0.911			8
	O15858	0.943			7
	Q6VVE5	0.927			7
Plasmodium yoelii yoelii [25]	Q7RA67	0.555	18	2	0
Coccidioides posadasii [26]	A1KXE9	0.887	185	40	6

The proteomes of the respective species were taken from Swiss-Prot [11] and used as input. Some of the proteins had ACC values out of the range of the training set and were considered as outliers. Eighteen antigens out of a total of 26 positive proteins were predicted to be immunogens (*sensitivity* 69%). According to the predictive score, the true positives were ranked in the first 25% of their proteomes.

VaxiJen 2.0 – Server for *in silico* Prediction of Immunoprotective Proteins

Predictive models for parasite and fungal immunoprotective proteins derived in the present study were incorporated into the VaxiJen server. VaxiJen is a server for *in silico* prediction of proteins with immunoprotective ability. It is freely accessible at the following URL: http://www.jenner.ac.uk/VaxiJen. The first version of the server included models for prediction of three types of antigen: bacterial, viral and tumour immunogens. The new version also includes models for parasite and fungal protective antigens. Proteins sequences can be submitted as single proteins or uploaded as a multiple sequence file in fasta format. Output of ACC coefficients and sequence are optional. The results page lists the selected type of antigen, the protein sequence, its probability score, and a statement of protective antigen or non-antigen, as defined by a threshold. Recommended thresholds are given according to the highest *accuracy* of the models.

DISCUSSION

Sequence alignment-based bioinformatics tools cannot identify novel antigenic proteins which lack sequence similarity to known antigens. Recently, we devised an alignment-free method for predicting immunoprotective proteins, based on ACC preprocessing of protein sequences. We developed this technology and applied it to bacterial, viral and tumour antigens [9,10]. The obtained results were encouraging and we have now extended our approach to include parasite and fungal immunoprotective antigens. In the present study, an immunoprotective antigen is either a protein (or part thereof) or its corresponding DNA known to be protective in an appropriate animal model following immunization. The models derived here are strictly kingdom specific and had good predictive ability, showing 69% *sensitivity* in the external validation and ranking the true immunoprotective proteins in the first 25% of their proteomes. The models discriminate between immunoprotective antigens and non-antigens without considering explicitly the presence or absence of T-cell or/and B-cell epitopes.

The usage of *z* descriptors to encode the protein sequence condenses physico-chemical property information along the protein sequence. Amino acids with similar properties have similar values for their *z* descriptors and are considered bioisosteric. The term "bioisosteric equivalent" refers to compounds or groups that possess near equal molecular shapes and volumes, approximately the same distribution of electrons, and which exhibit similar physical and biological properties (for example aspartic and glutamic acid or lysine and arginine). ACC transformations remove the influence of sequence length and account for the covariance of similar amino acids. Antigenicity and immunogenicity are not simple linear properties, and ACC preprocessing of physicochemical properties reflects adequately the discrimination between antigens and non-antigens.

The models derived in the present study were implemented in the second version of the server VaxiJen. VaxiJen 2.0 is freely accessible *via* http://www.jenner.ac.uk/VaxiJen. It assesses a protein's ability to be a protective antigen. The server can deal with either single proteins or whole proteomes. A higher score of a protein refers to a higher probability for protective ability. The final set of immunoprotective proteins represents no more than 25% of the species proteome and contains 69% of the true immunogens, as indicated by the external validation. The server can be used singly or in combination with alignment-dependent prediction and/or methods for predicting subcellular location. VaxiJen and the methods it implements represent a key contribution to *in silico* vaccinology.

ACKNOWLEDGEMENTS

The present study was supported by grants from the Medicinal Research Council of the Medical University of Sofia and the Ministry of Education and Science, Bulgaria. One of us (DRF) acknowledges funding from the Jenner Foundation.

REFERENCES

[1] State of the World's Vaccines and Immunization. WHO, **2002**, http://www.who.int/vaccines-documents/.

[2] Rappuoli, R. *Vaccine*, **2001**, *19*, 2688.

[3] Jenkins, M.C. *Vet. Parasitol.*, **2001**, *101*, 291.

[4] Mora, M.; Veggi, D.; Santini, L.; Pizza, M.; Rappuoli, R. *Drug Discov. Today*, **2003**, *8*, 459.

[5] Altschul, S.F.; Gish, W.; Miller, W.; Myers, E.W.; Lipman, D.J. *J. Mol. Biol.*, **1990**, *215*, 403.

[6] Nakai, K.; Kanehisa, M. *Proteins*, **1991**, *11*, 95.

[7] Bendtsen, J.D.; Nielsen, H.; von Heijne, G.; Brunak, S. *J. Mol. Biol.*, **2004**, *340*, 783.

[8] Vivona, S.; Gardy, J.L.; Ramachandran, S.; Brinkman, F.S.; Raghava, G.P.; Flower, D.R.; Filippini, F. *Trends Biotechnol.*, **2008**, *26*, 190.

[9] Doytchinova, I.A.; Flower, D.R. *Vaccine*, **2007**, *25*, 856.

[10] Doytchinova, I.A.; Flower, D.R. *BMC Bioinform.*, **2007**, *8*, 4.

[11] UniProt Knowledgebase of the ExPASy Proteomics Server [http://ca.expasy.org/sprot/].

[12] Hellberg, S.; Sjöström, M.; Skagerberg, B.; Wold, S. *J. Med. Chem.*, **1987**, *30*, 1126.

[13] Wold, S.; Jonsson, J.; Sjöström, M.; Sandberg, M.; Rännar, S. *Anal. Chim. Acta*, **1993**, *277*, 239.

[14] SIMCA 8.0. Umetrics UK Ltd, Wokingham Road, RG42 1PL, Bracknell, UK.

[15] Bradley, A.P. *Pattern Recognit.*, **1997**, *30*, 1145.

[16] Verma, S.K.; Bansal, I.; Vedi, S.; Saxena, J.K.; Katoch, V.M.; Bhattacharya, S.M. *Parasitol. Res.*, **2008**, *102*, 481.

[17] Acosta, D.; Cancela, M.; Piacenza, L.; Roche, L.; Carmona, C.; Tort, J.F. *Mol. Biochem. Parasitol.*, **2008**, *158*, 52.

[18] Goto, Y.; Bogatzki, L.Y.; Bertholet, S.; Coler, R.N.; Reed, S.G. *Vaccine*, **2007**, *25*, 7450.

[19] Coler, R.N.; Goto, Y.; Bogatzki, L.; Raman, V.; Reed, S.G. *Infect. Immun.*, **2007**, *75*, 4648.

[20] Uribe, N.; Muro, A.; Vieira, C.; Lopez-Aban, J.; del Olmo, E.; Suarez, L.; Martinez-Fernandez, A.R.; Siles-Lucas, M. *J. Parasitol.*, **2007**, *93*, 964.

[21] Tang, L.; Zhou, Z.; Chen, Y.; Luo, Y.; Wang, L.; Chen, L.; Huang, F.; Zeng, X.; Yi, X. *Cell Mol. Immunol.*, **2007**, *4920*, 153.

[22] Lopera-Mesa, T.M.; Kushwaha, A.; Mohmmed, A.; Chauhan, V.S. *Vaccine*, **2008**, *26*, 1335.

[23] Roussihon, C.; Oeuvray, C.; Mueller-Graf, C.; Tall, A.; Roqier, C.; Trape, J.F.; Theisen, M.; Balde, A.; Perignon, J.L.; Druilhe, P. *PLoS Med.*, **2007**, *4*, e320.

[24] Kar, P.; Dash, A.P.; Supakar, P.C. *Mol. Biochem. Parasitol.*, **2007**, *155*, 156.

[25] Bayele, H.K.; Brown, K.N. *FEMS Immunol. Med. Microbiol.*, **2007**, *50*, 389.

[26] Herr, R.A.; Hung, C.Y.; Cole, G.T. *Infect. Immun.*, **2007**, *75*, 5777.

Received: March 25, 2008 Revised: June 06, 2008 Accepted: June 10, 2008