BIOINFORMATICS ORIGINAL PAPER

Vol. 21 no. 4 2005, pages 483–491
doi:10.1093/bioinformatics/bti028

SPAAN: a software program for prediction of adhesins and adhesin-like proteins using neural networks

*Gaurav Sachdeva†, Kaushal Kumar†, Preti Jain† and Srinivasan Ramachandran**

G.N. Ramachandran Knowledge Center for Genome Informatics, Institute of Genomics and Integrative Biology, Mall Road, Delhi 110 007, India

Received on February 24, 2004; revised on August 20, 2004; accepted on August 31, 2004
Advance Access publication September 16, 2004

ABSTRACT

Motivation: The adhesion of microbial pathogens to host cells is mediated by adhesins. Experimental methods used for characterizing adhesins are time-consuming and demand large resources. The availability of specialized software can rapidly aid experimenters in simplifying this problem. We have employed 105 compositional properties and artificial neural networks to develop SPAAN, which predicts the probability of a protein being an adhesin (P_{ad}).

Results: SPAAN had optimal sensitivity of 89% and specificity of 100% on a defined test set and could identify 97.4% of known adhesins at high P_{ad} value from a wide range of bacteria. Furthermore, SPAAN facilitated improved annotation of several proteins as adhesins. Novel adhesins were identified in 17 pathogenic organisms causing diseases in humans and plants. In the severe acute respiratory syndrome (SARS) associated human corona virus, the spike glycoprotein and nsps (nsp2, nsp5, nsp6 and nsp7) were identified as having adhesin-like characteristics. These results offer new lead for rapid experimental testing.

Availability: SPAAN is freely available through ftp://203.195.151.45

Contact: ramu@igib.res.in

INTRODUCTION

Microbial pathogens encode adhesins that mediate their adherence to host cell surface receptors, membranes or the extracellular matrix for successful colonization. Investigations into this primary event of host–pathogen interaction have revealed a wide array of adhesins in a variety of pathogenic microbes (Finlay and Falkow, 1997). New approaches to vaccine development focus on targeting adhesins to abrogate the colonization process (Wizemann *et al*., 1999). However, the specific roles of particular adhesins in several pathogens remain to be elucidated.

One of the best-understood mechanisms of bacterial adherence is the attachment mediated by pili or fimbriae. The well-studied adhesins in this category are FimH and PapG adhesins of *Escherichia coli* (Hahn *et al*., 2002) and the Type IV pili adhesins in *Pseudomonas aeruginosa*, Neisseria, Moraxella, enteropathogenic *E.coli* and *Vibrio cholerae* (Strom and Lory, 1993). Several adhesins from other commonly known bacterial pathogens include MrkD protein of *Kleibsiella pneumoniae* (Gerlach *et al*., 1989), Hia of *Haemophilus influenzae* (Barenkamp and St Geme, 1996) and many others (for further details see http://www.igib.res.in/data/seepath/spaan_data.html).

Several vaccine formulations either currently approved or being evaluated use adhesins as immunizing agents. Examples include filamentous hemagglutinin and pertactin proteins against *Bordetella pertussis* (Halperin *et al*., 2003), FimH against pathogenic *E.coli* (Langermann *et al*., 2000), PsaA against pneumococcal disease (Rapola *et al*., 2003), outer membrane vesicle preparations including BabA adhesin against *Helicobacter pylori* infections (Prinz *et al*., 2003) and a synthetic peptide anti-adhesin vaccine against *P.aeruginosa* infections (Cachia and Hodges, 2003).

Experimental identification of adhesins is an arduous task. Computational methods such as homology search can aid in identification, but this procedure suffers from limitations when the homologues are not characterized. Sequence analysis based on compositional properties provides relief to this problem. Amino acid composition is a fundamental attribute of a protein and it has significant correlation to its location, function, folding type, shape and *in vivo* stability (Nakashima and Nishikawa, 1994; Nandi *et al*., 2003). Recently, compositional properties have been applied to the problems as diverse as the prediction of functional roles (Hobohm and Sander, 1995), protein secondary structures (Rost and Sander, 1993), secretory proteins

*To whom correspondence should be addressed.

†The authors wish it to be known that, in their opinion, the first three authors should be regarded as joint First Authors.

and apicoplast-targeted proteins in *Plasmodium falciparum* (Schneider, 1999; Zuegge *et al*., 2001).

We report a non-homology method using 105 compositional properties combined with artificial neural networks (ANNs) to identify adhesins and adhesin-like proteins in species belonging to a wide phylogenetic spectrum.

SYSTEMS AND METHODS

The five attributes

Amino acid frequencies Amino acid frequency f_i = (counts of i-th amino acid in the sequence)/l, where $i = 1, \ldots, 20$ and l is the length of the protein.

Multiplet frequencies Multiplets are defined as homopolymeric stretches $(X)_n$ where X is the amino acid and n (integer) $\geqslant 2$ (Brendel *et al*., 1992). After identifying all the multiplets, the frequencies of the amino acids in the multiplets were computed as follows:

$f_i(m)$ = (counts of i-th amino acid occurring as multiplet)/l.

Dipeptide frequencies The frequency of a dipeptide (i, j) f_{ij} = (counts of ij-th dipeptide)/(total dipeptide counts), where $i, j = 1$–20.

The theoretical number of possible dipeptides is 400. The recommended ratio for the number of input vectors to the number of weight connections is $\sim$2 to avoid overfitting (Andrea and Kalayeh, 1991). Therefore, we used top 20 dipeptides (when arranged in the ascending order of the P-values assessed using t-test) whose frequencies in the adhesin dataset were significantly different from that in the non-adhesin dataset (single-letter code): NG, RE, TN, NT, GT, TT, DE, ER, RR, RK, RI, AT, TS, IV, SG, GS, TG, GN, VI and HR.

Charge composition The frequency of charged amino acids (R, K, E and D considering the ionization properties of the side chains at pH 7.2) is given by f_c = (counts of charged amino acids)/l Furthermore, information on the characteristics of the distribution of the charged amino acids in a given protein sequence was obtained by computing the moments of the positions of the occurrences of the charged amino acids.

The general expression to compute moments of a given order; say 'r' is

$$M_r = r\text{-th order moment of the positions of charged amino acids} = \sum \frac{(X_i - X_m)^r}{N},$$

where, X_m is the mean of all positions of charged amino acids, $X_m = \sum_{i=1}^{N} X_i/N$; X_i is the position of i-th charged amino acid; and N is the number of charged amino acids in the sequence.

The frequency of charged amino acids, the length of the protein and the moments of order from of 2 to 19 were used to train the ANN constituting a total of 20 inputs. Moments of order >19 were not useful in further enhancing the performance.

Hydrophobic composition The amino acids were classified into five groups based on their hydrophobicity scores: (-8 for K, E, D and R), (-4 for S, T, N and Q), (-2 for P and H), ($+1$ for A, G, Y, C and W) and ($+2$ for L, V, I, F and M) (Brendel *et al*., 1992).

The inputs fed into the neural network for each group are as follows:

(1) f_i = (counts of i-th group)/(total counts in the protein), where $i = 1$–5.

(2) m_{ji} = j-th order moment of positions of amino acids in i-th group, where $j = 2$–5.

A total of 25 inputs representing the hydrophobic composition of a protein were fed to the neural network.

Taken together, a total of 105 compositional properties in the five modules were used to predict the adhesin-like characteristics of a given protein sequence.

Database construction

Adhesins Protein sequences were retrieved from http://www.ncbi.nlm.nih.gov using the keyword 'adhesin'. Furthermore, proteins containing the following keywords were removed from the primary retrieval: 'transport', 'pyrophosphatase', 'peroxidase', 'myosin', 'chaperone', 'hydrolase', 'gene product', 'accessory', 'regulatory', 'patent', 'permease', 'hypothetical', 'keratin', 'agrobacterium', 'intimin', 'ORFA', 'ATP binding', 'tRNA', 'deiminase', 'metalloproteinase', 'cofactor', 'amylase', 'methylase', 'unknown', 'ribosomal', 'alternative start', 'submitter believes' and 'phospholipase'. The remaining sequences in the adhesin database were manually curated to generate a set of well-annotated proteins many of which have been verified experimentally.

Non-adhesins The rationale we used here was to collect sequences of enzymes and other proteins that function within the cell. They probably have remote possibility of functioning as adhesins and would differ in compositional characteristics (Nakashima and Nishikawa, 1994). The keywords used were 'dehydratase', 'dehydrogenase', 'ribosomal protein', 'kinase', 'polymerase', 'acyl-CoA synthase', 'decarboxylase', and 'hydrolase'. Since effective implementation of the algorithm requires that the sizes of the two databases to be similar, we selected sequences from *Methanococcus jannaschii, E.coli* and *Saccharomyces cerevisiae* as representatives of the three primary kingdoms of life: Archaea, Eubacteria and Eukarya. This selection offers a diverse set for obtaining a broad range of limits for the detection of non-adhesins. In the subsequent step, 'hypothetical',

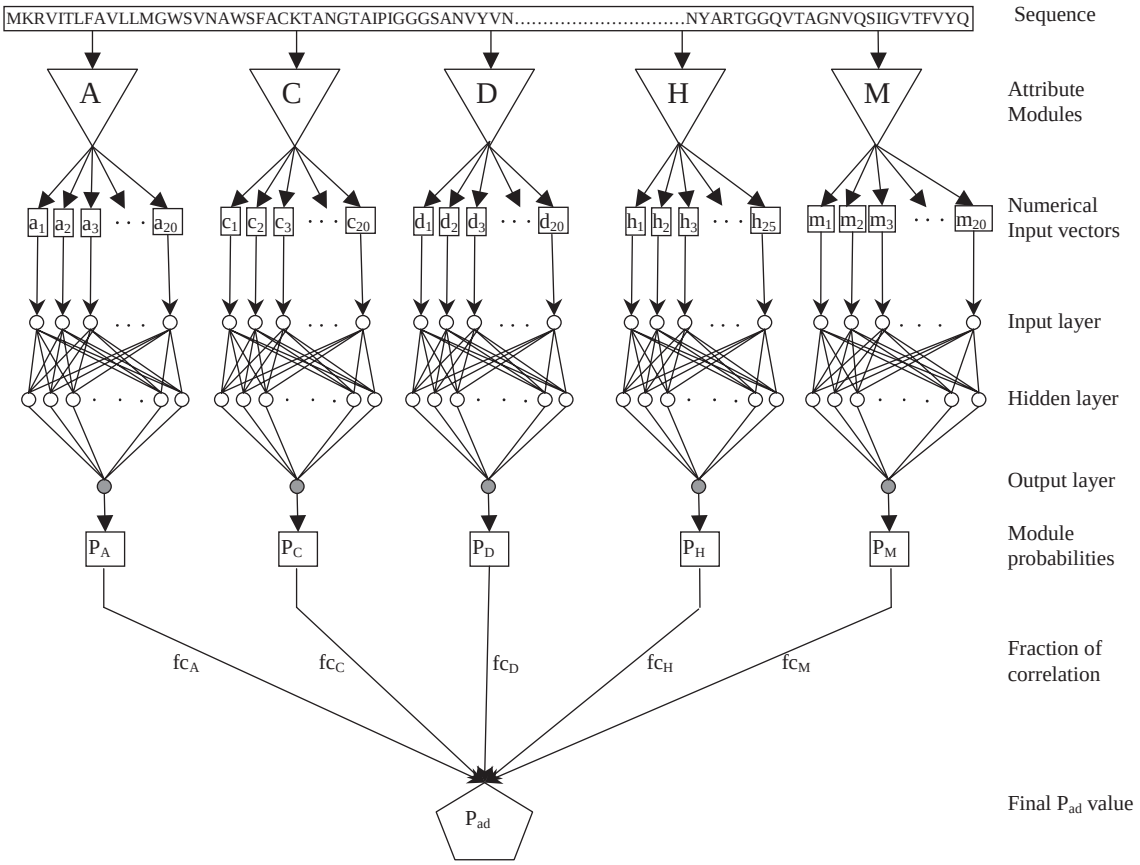

Fig. 1. The neural network architecture of SPAAN. A given protein sequence was processed through five modules, A, C, D, H and M, to quantify the five types of compositional attributes. A, Amino acid frequencies; C, Charge composition; D, Dipeptide frequencies; H, Hydrophobic composition; and M, Multiplet frequencies. The sequence shown is part of the FimH precursor (gi 5524634) of *E.coli*. The direction of arrows show data flow.

'transport', 'unknown' and 'membrane' protein sequences were removed.

Eliminating redundant entries We used CLUSTALW (Thompson *et al*., 1994) to analyze sequence similarities between the sequences in pairwise comparisons. Only one sequence entry was retained among pairs that had a CLUSTALW score of 100. Partial sequence entries were also removed. The total number of adhesins was 469 and the total number of non-adhesin proteins from *E.coli* was 282, *M.jannaschii* was 162 and *S.cerevisiae* was 259 which summed to 703 entries.

Neural network

The feed forward error back propagation neural network algorithm was used. The program was downloaded from the website http://www.cs.colostate.edu/~anderson a gift from Charles W. Anderson (Department of Computer Science, Colorado State University, Fort Collins, CO, e-mail: anderson@cs.colostate.edu)

ALGORITHM

Neural network architecture

The neural network used here has a multilayer feed forward topology. It consists of an input layer, a hidden layer and an output layer. This is a 'fully-connected' neural network where each neuron i is connected to each neuron j of the next layer (Fig. 1). The weight connections are denoted by w_{ij}. The state I_i of each neuron in the input layer is assigned directly from the input data, whereas the states of hidden layer neurons are computed from the states of input layer neurons using the sigmoid function,

$$h_j = 1 \Big/ \left[1 + \exp - \left(w_{j0} + \sum w_{ij} I_i\right)\right],$$

where w_{j0} is the bias weight. The back propagation algorithm was used to minimize the differences between the computed output and the target value. The target value for adhesins was set as '1' and for non-adhesins it was set as '0'.

In the initial optimization experiments, a training set and a validate set were used. The training set had 367 adhesins and

580 non-adhesins. The validate set had 102 adhesins and 123 non-adhesins. A total of 10 000 cycles (epochs) of training iterations were performed. Subsequently, the best epoch with minimum error on the validate set was identified and the corresponding weight matrix was used for the prediction.

Five networks were prepared, one for each attribute (Fig. 1). The number of neurons in the input layer was equal to the number of input data points for each attribute. The optimal number of neurons in the hidden layer was determined through experimentation for minimizing the error at the best epoch for each network individually. An upper limit for the total number of weight connections was set to half of the total number of input vectors to avoid overfitting as suggested previously (Andrea and Kalayeh, 1991). The final number of neurons in the hidden layer for each module was Amino acid frequencies: 30, Multiplets frequencies: 28, Dipeptide frequencies: 28, Charge composition: 30 and Hydrophobic composition: 30. During predictions, the network is fed with new data from the sequences that were not part of the training set.

Probability of being an adhesin, the P_{ad} value

Query proteins were processed modularly through the networks trained for each attribute. Thus, five probability outputs for each sequence were obtained. Final prediction was computed using the following expression, which is a weighted linear sum of the probabilities from five modules:

$$P_{ad} = \frac{(P_A * fc_A + P_C * fc_C + P_D * fc_D + P_H * fc_H + P_M * fc_M),}{(fc_A + fc_C + fc_D + fc_H + fc_M)}$$

where fc_i is the fraction of correlation of i-th module of the trained neural network, where $i = A$ (Amino acid frequencies), C (Charge composition), D (Dipeptide frequencies), H (Hydrophobic composition) or M (Multiplet frequencies). The fractions of correlation fc_i represent the fractions of total entries that were predicted correctly ($P_{i,\text{adhesin}} > 0.5$ and $P_{i,\text{non-adhesin}} < 0.5$) by the trained network on the validate set (Charles W. Anderson, http://www.cs.colostate.edu/~anderson). $fc_A = 0.84$, $fc_C = 0.71$, $fc_D = 0.84$, $fc_H = 0.79$, $fc_M = 0.83$.

Matthew's correlation coefficient for assessing the performance of SPAAN

The Matthew's correlation coefficient (Mcc) (Matthews, 1975) is defined as follows:

$$\text{Mcc} = \frac{(\text{TP} * \text{TN}) - (\text{FP} * \text{FN})}{\sqrt{(\text{TN} + \text{FN})(\text{TN} + \text{FP})(\text{TP} + \text{FN})(\text{TP} + \text{FP})}},$$

where TP stands for true positives, TN the true negatives, FP the false positives and FN the false negatives.

Here TPs are adhesins and TNs are non-adhesins. Adhesins with P_{ad} value above a chosen threshold are TPs, whereas known non-adhesins with P_{ad} value below the chosen threshold are TNs. The sensitivity, Sn, is given by (TP/(TP + FN)) and specificity, Sp, is given by (TP/(TP + FP)).

SPECIFICATIONS

Computer programs used to compute individual compositional attributes were written in C and executed on a PC with operating system Red Hat Linux version 7.3 or 8.0.

Sequence inputs

SPAAN accepts input sequence files in the FASTA format. Multiple sequences can be present in one file. Protein sequences with ambiguous amino acids and/or of length <50 amino acids were filtered out. Amino acids must use the single-letter code according to the IUPAC-IUB nomenclature system.

RESULTS AND DISCUSSION

Sensitivity, specificity and correlation coefficient

In designing SPAAN, we developed a non-homology, compositional property based method to predict adhesins and adhesin-like proteins solely from the sequence data. To assess the performance of SPAAN, we prepared a test set of 37 well-annotated adhesins and 37 non-adhesins that were not part of the training set. The results are shown in Figure 2. It is apparent that SPAAN could identify 89% of known adhesins with 100% specificity when examined at $P_{ad} \geq 0.51$. At $P_{ad} \geq 0.51$, the Mcc (Matthews, 1975) was observed to be highest (0.94). We observed that the combination of five modules provided the best results.

Assessment of the performance in individual modules showed that they performed poorly when compared with the combination of modules. The performances of individual modules were as follows: Charge composition ($P_c = 0.55$, Mcc = 0.658, Sn = 0.756 and Sp = 0.848), Dipeptide frequencies ($P_D = 0.54$, Mcc = 0.84, Sn = 0.86 and Sp = 0.94), Hydrophobic composition ($P_H = 0.61$, Mcc = 0.63, Sn = 0.54 and Sp = 0.9) and Multiplet frequencies ($P_H = 0.58$, Mcc = 0.77, Sn = 0.81 and Sp = 0.9). Performance of the Amino acid frequencies' module could not be assessed unambiguously because the Mcc was nearly flat over a broad range. These observations suggest that it would be fruitful to include multiple modules for obtaining high-quality predictions and are consistent with the experience of Hobohm and Sander (1995).

SPAAN predicts experimentally characterized adhesins with high P_{ad} value

Considering the small size of test set, we examined the general applicability of SPAAN by analyzing several well-characterized adhesins from a wide range of pathogens causing a variety of diseases. The results on 194 adhesins with binding specificity to a wide range of host receptors are displayed in Table 1 (for further details see http://www.igib.res.in/data/seepath/spaan_data.html). It is apparent that except two FimH proteins of *E.coli*, pertactin of *B.pertussis*, protein F of *Streptococcus pyogenes* and PspC

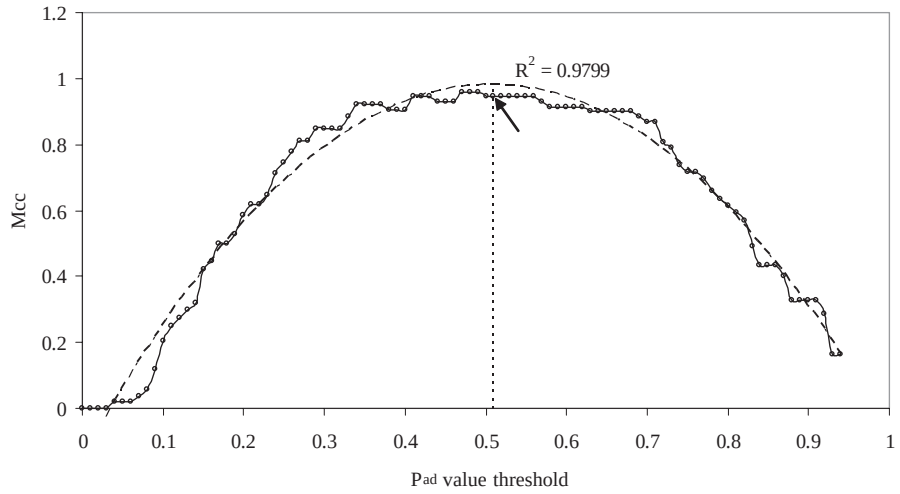

Fig. 2. Performance of SPAAN. A theoretical polynomial curve of second-order (dashed line) was fitted to the observed curve (smooth line) with a Karl–Pearson correlation coefficient $R^2 = 0.9799$. The maximum point of the theoretical curve was chosen as a reference (vertical dotted line) to identify the maximum Mcc on the observed curve (shown by an arrow). Note that the Mcc does not drop down to the x-axis because the highest P_{ad} value attained using adhesins was 0.939 in comparison with the theoretical attainable limit of 1.0.

Table 1. Prediction of well-characterized adhesins from various bacterial pathogens using SPAAN

Species	Disease caused	Adhesin[a]	Host ligand	P_{ad} value[b] (range)	Reference
E.coli	Diarrhoea	PapG (27)	α-D-gal(1–4) β-D-Gal-containing receptors	0.84–0.76	Bock *et al*. (1985)
		SfaS (5)	alpha-sialyl-beta-2,3-b-galactose	0.94–0.94	Moch *et al*. (1987)
		FimH (63)	D-mannosides	0.96–0.23[c]	Hahn *et al*. (2002)
		Intimin (12)	Tyrosine-phosphorylated form of host cell receptor Hp90	0.95–0.78	Rosenshine *et al*. (1996)
		PrsG (5)	Gal(alpha1–4)Gal	0.86–0.85	Johnson *et al*. (1997)
Non-typeable *H.influenzae*	Influenza	HMW1, HMW2	Human epithelial cells	0.97	St Geme (1996)
		Hia (8)	Human conjuctival cells	0.93–0.90	
B.pertussis	Whooping cough	FHA	Sulfated sugars on cell-surface glycoconjugates	0.85	Brennan and Shahin (1996)
		Pertactin	Integrins	0.43	Brennan and Shahin (1996)
Yersinia enterocolitica	Enterocolitis	YadA (5)	β_1 integrins	0.88–0.79	Schulze-Koops *et al*. (1993)
Enterococcus faecalis	Empyma in patients with liver disease	EfaA	Unknown	0.83	Finlay and Falkow (1997)
H.pylori	Peptic ulcers	BabA (17)	Difucosylated Lewis blood group antigen	0.87–0.68	Prinz *et al*. (2003)

[a]The number of sequences analyzed from different strains and homologs from related species are shown in parentheses.
[b]Rounded off to the second decimal.
[c]Out of 63 FimH proteins, 54 were from *E.coli*, 6 from *Shigella flexineri*, 2 from *Salmonella enterica* and 1 from *Salmonella typhimurium*. Except two FimH proteins, the rest had $P_{ad} \geq 0.51$. The two exceptions (gi numbers: 5524636 and 1778448) were from *E.coli*. The gi: 5524636 protein is annotated as a FimH precursor but is much shorter (129 amino acids) than other members of the family. The gi: 1778448 protein is a *S.typhimurium* homolog in *E.coli*.

of *Streptococcus pneumoniae*, the rest 189 adhesins had $P_{ad} \geq 0.51$ indicating an overall sensitivity of 97.4%. These results demonstrate the general applicability of SPAAN.

SPAAN is a non-homology method based on sequence properties

To examine the non-homology character of SPAAN, we prepared a dataset of 130 adhesins that did not have any protein pairs with CLUSTALW score of 100. Equal number of non-adhesins was selected with the same criterion. A histogram plot of adhesins and non-adhesins in the various ranges of P_{ad} values is displayed in Figure 3a. It is evident that SPAAN is capable of segregating the adhesins and non-adhesins into two distinct cohesive groups. Most of the adhesins (96%) have $P_{ad} \geq 0.51$ whereas all the non-adhesins (100%) have $P_{ad} < 0.51$.

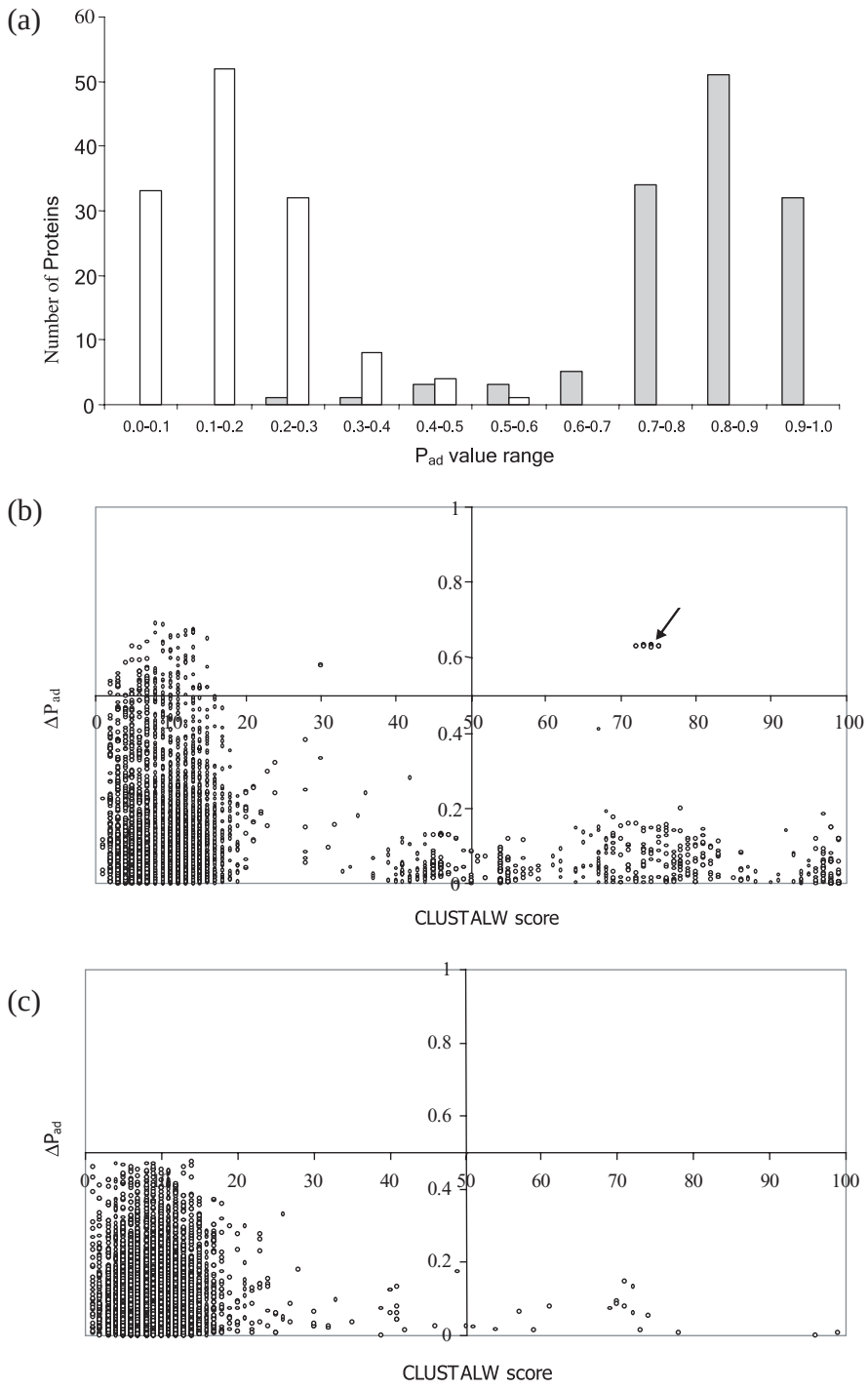

Fig. 3. SPAAN is a non-homology-based software program. A total of 130 adhesins and 130 non-adhesins were analyzed to assess whether the predictive power of SPAAN could be influenced by sequence relationships. (**a**) Histogram plots of the number of proteins in the various P_{ad} value ranges are shown. Shaded bars represent adhesins and open bars represent non-adhesins. Note the ability of SPAAN to segregate adhesins and non-adhesins into two distinct cohesive groups. (**b**) Pairwise sequence relationships among the adhesins were determined using CLUSTALW and plotted on *x*-axis. Higher CLUSTALW scores indicate similar pairs. The corresponding differences in P_{ad} values in the same protein pair was plotted on the *y*-axis. Each point in the diagram represents a pair. Arrow points to protein pairs of the FimH family with high ΔP_{ad} values in spite of high similarity. Since one of the FimH proteins (gi: 5524636) had very low P_{ad} value, all pairs with this false negative protein show high ΔP_{ad} values. The protein (gi: 5524636) is of much shorter length compared with other members of the same family. (**c**) Plot for non-adhesins. Data were plotted in the four-quadrant format to enhance clarity.

Table 2. Analysis of predictions carried out using SPAAN on genome scans of a few selected pathogenic organisms

Organism	Disease caused	Total no. of proteins analyzed	No. of these supported by complementary evidence CDD/BLASTP[a]/PubMed	No. of these supported by complementary evidence BETAWRAP	No. of adhesin-like proteins	No. of false positives
E.coli O157:H7	Diarrhoea	50	37[b]	33	12	1
H.pylori	Peptic ulcers	50	25[c]	36	24	1
Listeria monocytogenes	Listeriosis	50	23[d]	39	24	3
S.pneumoniae R6	Bacterial pneumonia	40	13[e]	12	23	4
Mycobacterium tuberculosis H37Rv	Tuberculosis	50	—	32	50[f]	—
SARS-associated corona virus	SARS	5	—	—	5[g]	—

[a]The top-scoring similar sequences with $E < 0.001$ were only considered for assessing sequence-based relationships. The low complexity filter was 'off'.
[b]Includes Fimbrial adhesins (nine proteins), AidA-I, gamma intimin, hemagglutinin, translocated intimin receptor, putative tail fiber protein and putative major tail protein.
[c]Includes putative vacuolating cytotoxin (VacA) (autotransporter adhesin-like), outer membrane protein (adhesin; 2 proteins), outer membrane protein (porin; 3 proteins), flagellin A and outer membrane proteins (13 proteins).
[d]Includes internalin A (mediates adhesion and invasion), other internalins (seven proteins), peptiglycan-linked protein (similar to autotransporter adhesin, eight proteins), autolysin (amidase, presumably involved in adhesion, three proteins), flagellar hook protein may be involved in adhesion, cell surface protein (adhesin by BLASTP analysis).
[e]Includes PsaA (pneumococcal surface adhesin A), pspA similar to CbpA (choline binding protein A), CbpA, CbpD, CbpE, CbF and Cbp (two proteins) all similar to CbpA, two hypothetical proteins with low-level similarity to CbpA, autolysin.
[f]PE, PE_PGRS (35 proteins), PPE (12 proteins), two hypothetical proteins with no similarity to either PE or PE_PGRS or PPE proteins.
[g]These proteins were the spike glycoprotein with antigenic properties, and nsp2, nsp5, nsp6 and nsp7.

We computed the pairwise sequence similarities using CLUSTALW (Thompson *et al*., 1994) and compared with the differences between the P_{ad} values in the pair (denoted by ΔP_{ad}). The relationships for adhesins and non-adhesins are shown in Figures 3b and c, respectively. In both adhesins and non-adhesins, ΔP_{ad} values are uniformly low and appear nearly independent of sequence relationship. Furthermore, among the protein pairs with score <20, 82% of adhesin pairs and 86% of non-adhesin pairs had $\Delta P_{ad} < 0.2$. These data reinforce the non-homology character of SPAAN.

Application of SPAAN to whole genomes

The results of the genome scan for selected pathogens of humans and plants are displayed in Table 2 (for detailed data description, see online Supplementary Table at http://www.igib.res.in/data/seepath/spaan_data.html). We used a stringent criterion of $P_{ad} > 0.7$ on the basis of the results shown in Figure 3a to reduce the detection of false positives. Subsequently, we restricted our analysis to a maximum of 50 top-scoring proteins. This serves as a good starting point to examine the performance of SPAAN and to identify top-scoring novel adhesins with high confidence. The experimentally characterized adhesins from a wide range of pathogens top the list in genome scans. Several of the predicted adhesins are supported by complementary evidence from Conserved Domain Database search (RPS-BLASTP), BLASTP and the beta helix predictor BETAWRAP (Marchler-Bauer *et al*., 2002; Altschul *et al*., 1990; Bradley *et al*., 2001). About 30–78% of these predicted adhesins also contain beta helix motif. The beta helix motif was found to be associated with several adhesins, toxins, virulence factors and surface proteins (Bradley *et al*., 2001).

In addition, SPAAN guided the improved annotation of a number of adhesins by suggesting re-examination of these proteins using the most commonly used software listed above. It is also evident that the well-known adhesins in these organisms top the list of predictions using SPAAN (Table 2). Several proteins with high P_{ad} values were identified using SPAAN for which either limited or no complementary evidence exist. We have classified these proteins as 'adhesin-like'. Interestingly, several mycobacterial proteins, namely, 35 PE_PGRS proteins and 12 PPE proteins were identified with high P_{ad} value. SPAAN could identify these putative mycobacterial adhesins even though our training dataset was devoid of mycobacterial proteins. Indeed, experimental analysis has demonstrated that some of these proteins could mediate host–pathogen interactions (Brennan *et al*., 2001). These results demonstrate that SPAAN could overcome taxonomic limits and can be used for general purpose.

Although SPAAN was primarily trained on bacterial adhesins, we examined its ability to predict putative adhesins from eukaryotic systems. The criteria was relaxed by using the base threshold value of $P_{ad} \geq 0.51$ to scan the genome of SARS-associated human coronavirus. The spike glycoprotein, nsp2, nsp5, nsp6 and nsp7 were identified as adhesins. Spike glycoprotein has been implicated to play a role in viral entry and pathogenesis (Gallagher and Buchmeier, 2001). The role of nsp proteins in viral pathogenesis is not clear. Since SARS is an important public health problem, these results could rapidly aid experiments that characterize host–pathogen interactions.

A few false positives do appear in the list. A judicious approach for experimental characterization could be developed by considering the total number of proteins to be analyzed, prioritizing proteins with other complementary evidence while keeping the number of false positives as low as possible. In summary, SPAAN could serve as an useful guide to perform experimental characterization of proteins for adhesin-like properties.

ACKNOWLEDGEMENTS

We thank Prof. G. Padmanaban, Prof. Samir K. Brahmachari, Dr R. Sonti and A. Maharana for their useful discussions. S.R. thanks Council of Scientific and Industrial Research for a grant under the New Millennium Indian Technology Leadership Initiative (NMITLI) program.

REFERENCES

Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.

Andrea,T.A. and Kalayeh,H. (1991) Applications of neural networks in quantitative structure–activity relationships of dihydrofolate reductase inhibitors. *J. Med. Chem.*, **34**, 2824–2836.

Barenkamp,S.J. and St Geme,J.W.,III (1996) Identification of a second family of high-molecular-weight adhesion proteins expressed by non-typable *Haemophilus influenzae*. *Mol. Microbiol.*, **19**, 1215–1223.

Bock,K., Breimer,M.E., Brignole,A., Hansson,G.C., Karlsson,K.A., Larson,G., Leffler,H., Samuelsson,B.E., Stromberg,N., Eden,C.S. *et al*. (1985) Specificity of binding of a strain of uropathogenic *Escherichia coli* to Gal alpha 1–4Gal-containing glycosphingolipids. *J. Biol. Chem.*, **260**, 8545–8551.

Bradley,P., Cowen,L., Menke,M., King,J. and Berger,B. (2001) BETAWRAP: successful prediction of parallel beta-helices from primary sequence reveals an association with many microbial pathogens. *Proc. Natl Acad. Sci. USA*, **98**, 14819–14824.

Brendel,V., Bucher,P., Nourbakhsh,I.R., Blaisdell,B.E. and Karlin,S. (1992) Methods and algorithms for statistical analysis of protein sequences. *Proc. Natl Acad. Sci. USA*, **89**, 2002–2006.

Brennan,M.J. and Shahin,R.D. (1996) Pertussis antigens that abrogate bacterial adherence and elicit immunity. *Am. J. Respir. Crit. Care Med.*, **154**, S145–S149.

Brennan,M.J., Delogu,G., Chen,Y., Bardarov,S., Kriakov,J., Alavi,M. and Jacobs,W.R. (2001) Evidence that Mycobacterial PE_PGRS proteins are cell surface constituents that influence interactions with other cells. *Infect. Immun.*, **69**, 7326–7333.

Cachia,P.J. and Hodges,R.S. (2003) Synthetic peptide vaccine and antibody therapeutic development: prevention and treatment of *Pseudomonas aeruginosa*. *Biopolymers*, **71**, 141–168.

Finlay,B.B. and Falkow,S. (1997) Common themes in microbial pathogenicity revisited. *Microbiol. Mol. Biol. Rev.*, **61**, 136–169.

Gallagher,T.M. and Buchmeier,M.J. (2001) Coronavirus spike proteins in viral entry and pathogenesis. *Virology*, **279**, 371–374.

Gerlach,G.F., Clegg,S. and Allen,B.L. (1989) Identification and characterization of the genes encoding the type 3 and type 1 fimbrial adhesins of *Klebsiella pneumoniae*. *J. Bacteriol.*, **171**, 1262–1270.

Hahn,E., Wild,P., Hermanns,U., Sebbel,P., Glockshuber,R., Haner,M., Taschner,N., Burkhard,P., Aebi,U. and Muller,S.A. (2002) Exploring the 3D molecular architecture of *Escherichia coli* type 1 pili. *J. Mol. Biol.*, **323**, 845–857.

Halperin,S.A., Scheifele,D., Mills,E., Guasparini,R., Humphreys,G., Barreto,L. and Smith,B. (2003) Nature, evolution, and appraisal of adverse events and antibody response associated with the fifth consecutive dose of a five-component acellular pertussis-based combination vaccine. *Vaccine*, **21**, 2298–2306.

Hobohm,U. and Sander,C. (1995) A sequence property approach to searching protein databases. *J. Mol. Biol.*, **251**, 390–399.

Johnson,J.R., Russo,T.A., Scheutz,F., Brown,J.J., Zhang,L., Palin,K., Rode,C., Bloch,C., Marrs,C.F. and Foxman,B. (1997) Discovery of disseminated J96-like strains of uropathogenic *Escherichia coli* O4: H5 containing genes for both PapG (J96) (class I) and PrsG (J96) (class III) Gal(alpha1–4)Gal-binding adhesins. *J. Infect. Dis.*, **175**, 983–988.

Langermann,S., Mollby,R., Burlein,J.E., Palaszynski,S.R., Auguste,C.G., DeFusco,A., Strouse,R., Schenerman,M.A., Hultgren,S.J., Pinkner,J.S. *et al*. (2000) Vaccination with FimH adhesin protects cynomolgus monkeys from colonization and infection by uropathogenic *Escherichia coli*. *J. Infect. Dis.*, **181**, 774–778.

Marchler-Bauer,A., Panchenko,A.R., Shoemaker,B.A., Thiessen,P.A., Geer,L.Y. and Bryant,S.H. (2002) CDD: a database of conserved domain alignments with links to domain three-dimensional structure. *Nucleic Acids Res.*, **30**, 281–283.

Matthews,B.W. (1975) Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta*, **405**, 442–451.

Moch,T., Hoschutzky,H., Hacker,J., Kroncke,K.D. and Jann,K. (1987) Isolation and characterization of the alpha-sialyl-beta-2,3-galactosyl-specific adhesin. *Proc. Natl Acad. Sci. USA*, **84**, 3462–3466.

Nakashima,H. and Nishikawa,K. (1994) Discrimination of intracellular and extracellular proteins using amino acid composition and residue pair frequencies. *J. Mol. Biol.*, **238**, 54–61.

Nandi,T., Dash,D., Ghai,R., B-Rao,C., Kannan,K., Brahmachari,S.K., Ramakrishnan,C. and Ramachandran,S. (2003) A novel complexity measure for comparative analysis of protein sequences from complete genomes. *J. Biomol. Struct. Dyn.* **20**, 657–668.

Prinz,C., Hafsi,N. and Voland,P. (2003) *Helicobacter pylori* virulence factors and the host immune response: implications for therapeutic vaccination. *Trends Microbiol.*, **11**, 134–138.

Rapola,S., Jäntti,V., Eerola,M., Mäkelä,P.H., Käyhty,H. and Kilpi,T. (2003) Anti-PsaA and the risk of pneumococcal AOM and carriage. *Vaccine*, **21**, 3608–3613.

Rosenshine,I., Ruschkowski,S., Stein,M., Reinscheid,D.J., Mills,S.D. and Finlay,B.B. (1996) A pathogenic bacterium triggers epithelial signals to form a functional bacterial receptor that mediates actin pseudopod formation. *EMBO J.*, **15**, 2613–2624.

Rost,B. and Sander,C. (1993) Improved prediction of protein secondary structure by use of sequence profiles and neural networks. *Proc. Natl Acad. Sci. USA*, **90**, 7558–7562.

Schneider,G. (1999) How many potentially secreted proteins are contained in a bacterial genome? *Gene*, **237**, 113–121.

Schulze-Koops,H., Burkhardt,H., Heesemann,J., Kirsch,T., Swoboda,B., Bull,C., Goodman,S. and Emmrich,F. (1993) Outer membrane protein YadA of enteropathogenic yersiniae mediates specific binding to cellular but not plasma fibronectin. *Infect. Immun.*, **61**, 2513–2519.

Strom,M.S. and Lory,S. (1993) Structure–function and biogenesis of the type IV pili. *Annu. Rev. Microbiol.*, **47**, 565–596.

St Geme,J.W.,III (1996) Progress towards a vaccine for nontypable *Haemophilus influenzae*. *Ann. Med.*, **28**, 31–37.

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

Wizemann,T.M., Adamou,J.E. and Langermann,S. (1999) Adhesins as targets for vaccine development. *Emerg. Infect. Dis.*, **5**, 395–403.

Zuegge,J., Ralph,S., Schmuker,M., McFadden,G.I. and Schneider,G. (2001) Deciphering apicoplast targeting signals—feature extraction from nuclear-encoded precursors of *Plasmodium falciparum* apicoplast proteins. *Gene*, **280**, 19–26.