Bioinformation by Biomedical Informatics Publishing Group *open access*

www.bioinformation.net **Prediction Model**

An SVM method using evolutionary information for the identification of allergenic proteins

Kandaswamy Krishna Kumar[1,*] and Prakash Shrikrishna Shelokar[1]

[1]Insilico Consulting, 402, Citi Centre, 39/2 Erandwane, Karve Road, Pune-411004, Maharashtra, India; Kandaswamy Krishna Kumar* - E-mail: biotechkk@gmail.com; * Corresponding author

received December 28, 2007; revised January 17, 2008; accepted January 19, 2008; published January 27, 2008

Abstract:
This study presents an allergenic protein prediction system that appears to be capable of producing high sensitivity and specificity. The proposed system is based on support vector machine (SVM) using evolutionary information in the form of an amino acid position specific scoring matrix (PSSM). The performance of this system is assessed by a 10-fold cross-validation experiment using a dataset consisting of 693 allergens and 1041 non-allergens obtained from Swiss-Prot and Structural Database of Allergenic Proteins (SDAP). The PSSM method produced an accuracy of 90.1% in comparison to the methods based on SVM using amino acid, dipeptide composition, pseudo (5-tier) amino acid composition that achieved an accuracy of 86.3, 86.5 and 82.1% respectively. The results show that evolutionary information can be useful to build more effective and efficient allergen prediction systems.

Keywords: allergenic proteins; evolutionary information; PSSM; amino-acid composition; dipeptide composition; SVM

Background:
A protein specific allergic reaction is the binding of IgE antibodies to an allergen, which result into common allergic reactions, such as, asthma, rhinitis, rhinoconjunctivitis, eczema, contact dermatitis, angioedema and abdominal pain. Different types of food, pollen or dust mites are common sources of allergens **[1]**. The introduction of many transgenic proteins into the food-chain and medical application has pressed for the development of different methods to safely conclude on potential protein allergenicity.

Prediction methods based on artificial intelligence and machine learning have gained popularity due to the available data on IgE epitopes, amino acid descriptors and several other variables. These applications include motif based approach using MEME/MAST **[2]**, *k*-nearest neighbor classifier **[3]**, similarity search against IgE epitopes, epitope profiles and structure profiles **[4-7]**. In this paper, a standard method has been developed for predicting allergens based on evolutionary information in the form of amino acid position specific scoring matrix (PSSM) **[8]** using support vector machine (SVM) **[4, 9, 10]**. The performance of the proposed system was assessed by a 10-fold cross-validation experiment on the dataset obtained from Swiss-Prot and SDAP database. For comparison with PSSM based SVM, we also developed methods based on SVM using amino acid, dipeptide composition and pseudo amino acid composition.

Methodology:

Training data

To develop our methodology, we obtained the data from different sources as:

Dataset 1

This dataset of 664 allergens was obtained from Li and colleagues **[6]**. From this dataset 68 sequences annotated with 'fragment' were removed. The resultant dataset of 596 sequences was employed to train our methodology.

Dataset 2

To validate our methodology, 97 allergen sequences were obtained from SDAP (Structural Database of Allergenic Proteins) **[7, 16]**.

Dataset 3

1041 non-allergen sequences were collected from Swiss-Prot. One of the criteria was organism: Lycopersicon (tomato), Apium (celery) or Pyrus (pear), commonly consumed commodities. Moreover, three clusters of exclusion criteria were applied: i) allergen or allergy (all text); ii) lipid-transfer protein, cupin, chitinase, profilin (all text); iii) fragment (all text). Most entries were originated from Lycopersicon.

Dataset 1 and Dataset 2 consist of allergen sequences while Dataset-3 consists of non-allergen sequences. Training dataset contains 1540 sequences with 596 allergens (positive sequences) and 944 non-allergens (negative

sequences) while test dataset contains a total of 194 sequences with all positive sequences (97) from Dataset 2 and equal number of negative sequences (97) from Dataset 3.

SVM binary classification

Recently SVM (Support Vector Machine) has shown many applications in the field of bioinformatics **[4, 9, 10]**. SVM is a supervised machine learning method which is based on the statistical learning theory **[12]**. When used as a binary classifier, an SVM will construct a hyperplane, which acts as the decision surface between the two classes. This is achieved by maximizing the margin of separation between the hyperplane and those points nearest to it. The details of the formulation and solution methodology of SVM for binary classification task can be found elsewhere **[12]**.

SVM software: LIBSVM

Simulations were preformed using LIBSVM version 2.81 (a freely available software package) **[17]**. The SVM training has been carried out by the optimization of the value of the regularization parameter and the value of RBF kernel parameter.

Input features

Amino acid composition

Amino acid composition is a fraction of each amino acid present in the protein sequence. If L is the length of protein and Qi is the frequency of occurrence of an amino acid i, then amino acid composition is $Ci = Qi/L$, where, i is any of the 20 amino acids.

Dipeptide composition

It transforms a protein into an input vector of 400 dimensions (20 by 20). If Qij be a fraction of a pair of amino acids (i, j = 1,...,20) and L be a total number of all possible dipeptides (L = 400) then the dipeptide composition is $Cij = Qij/L$, where i, j are any of the 20 amino acid residues.

Pseudo-amino acid composition

Pseudo amino acid composition is a representation of both the amino acid composition and sequence order effect **[13]**. Pseudo amino acid composition features were generated using PseAA web server **[18]**.

Position Specific Scoring Matrix (PSSM)

The PSSM for each query sequence was generated using three rounds of PSI-BLAST against a non-redundant protein database, with an E-value cut-off of 0.001 **[19]**. The PSSM provides a matrix of dimension L rows and 20 columns for a protein chain of L amino acid residues, where, 20 columns represent occurrence/substitution of each type of 20 amino acids. This PSSM matrix was further transformed into input vector of 400 dimensions using the methodology of Xie and colleagues **[14]**.

Performance evaluation

Different SVM models using PSSM, amino acid composition, dipeptide composition and pseudo amino acid composition were developed. In this study, allergen Dataset-1 was applied for training and Dataset-2 was employed for validation of different models. Several performance evaluation measures were employed as:

Cross validation experiments

A 10-fold cross-validation experiment **[15]** was carried out to evaluate the performance of the SVM models. The Dataset-1 was randomly divided into 10 subsets. The training and testing were carried out 10 times for each model using one distinct set for testing and the remaining nine for training. The performance of the model was reported as the average performance over 10 sets.

Re-substitution test

This test was applied to identify the self-consistency of models **[15]**. Training dataset itself was applied as the test set to verify the self-consistency of the model which was trained using 10-fold cross-validation process.

Validation test

Independent dataset consisting of all the 97 allergens from Dataset-2 and equal number of non-allergen sequences from Dataset-3 was used to evaluate the performance of different SVM models. A confusion matrix was employed to quantify the efficiency of classification between allergens from non-allergens using TP (True positive – known and predicted allergens), TN (True negative – known and predicted non allergens), FP (False positive – known non-allergens and predicted allergens) and FN (False negative – known allergens and predicted non allergens). We further define sensitivity (TP/ (TP+FN)), specificity (TN/ (TN+FP)), accuracy, positive predictive value (TP/ (TP+FP)), Negative Predictive value (TN/(TN+FN)), Matthews correlation coefficients (MCC) and F-measure 2*sensitivity*specificity)/ (sensitivity + specificity).

Discussion:

The aim of this study was to try evolutionary information using PSSM to build SVM models for allergen prediction. For comparison study, we also developed SVM models using amino acid composition, dipeptide composition and pseudo amino acid composition. Different SVM models were trained using 10-fold cross-validation process and tested using resubstitution test. The results are shown in Figure 1a and Figure 1b. It shows average accuracy in 10-fold cross validation experiments and resubstitution test.

SVM model using PSSM based input features achieved highest 10-fold cross-validation accuracy of 90.1% with resubstitution test accuracy of 98.7%. Other SVM models using pseudo amino acid, amino acid and dipeptide composition based input features achieved an accuracy of 82.1%, 86.3% and 86.5% with resubstitution test accuracy of 92.3%, 98.9% and 99.2% respectively. The hybrid

approach was tested by combining features from PSSM (400), amino acid composition (20) and pseudo amino acid composition (400). This approach produced a 10-fold cross validation accuracy of 91.5%. However, it has not shown a significant improvement in the performance as compared to that of PSSM based method that used input vector of 400 features.

Figure1b shows different performance measures evaluated on the independent dataset (194 sequences). The PSSM based input features applied by the SVM model has produced highest sensitivity of 92.8 with a positive predictive value and negative predictive value of 92.8 on the independent dataset. This approach also achieved the highest MCC of 0.856 and the F-measure of 0.93. The dipeptide and amino acid composition based SVM models predicted allergenicity of protein with a slightly lower sensitivity of about 88% and 86%, respectively. Both the models showed values of positive predictive value exceeding 90 %. The values for MCC obtained for these models were 0.8 and 0.82, respectively with F-measure of 0.89 and 0.90. The performance of the dipeptide based method was comparable to that of the amino acid composition-based method in terms of accuracy (86.5% vs. 86.3%), MCC and F-measure (Figure 1b), though di-peptides provide more information than amino acid composition. We examined our results and found that the frequency of occurrence of most dipeptides was low. The performance of the dipeptide composition-based approach was better than pseudo-amino acid composition-based approach. The most likely reason for low performance of the pseudo-amino acid composition-based approach may be that it considers only identical pairs of amino acids and ignores non-identical pairs. The dipeptide composition based approach considers all of the contiguous pairs of amino acids irrespective of identity.

The performance PSSM based SVM model was also compared to AlgPred **[4]** and WebAllergen **[5, 6]** models (see Figure 1c). We submitted an independent dataset of 97 allergens and equal number of non allergens to these models. Results show that the allergen prediction models based on PSSM using SVM appears to be capable of producing high positive and negative predictive value.

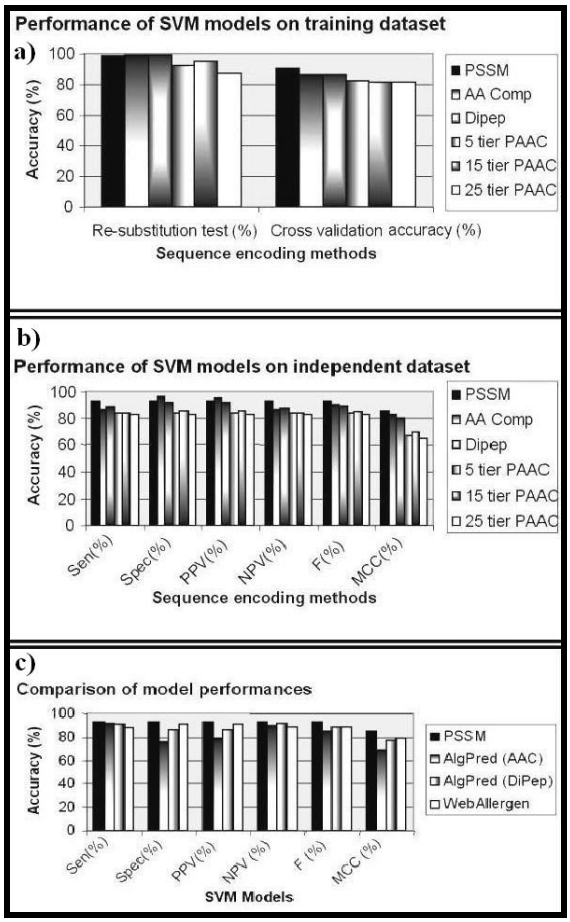

Figure 1: SVM model (a) performance in training set; (b) performance in independent set; (c) comparison with other models. AA comp/AAC – amino acid composition; Dipep – dipeptide composition; PAAC – pseudo amino acid composition; sen – sensitivity; spec – specificity; F – F-Measure.

Conclusion:
An SVM model using PSSM is developed to distinguish allergens and non-allergens. The results show that this approach has exhibited superior performance in comparison to the other approaches based on amino acid composition, dipeptide composition and pseudo amino acid composition. The results indicate that evolutionary information based input features can provide good amount of discriminative information, which may help to build more effective and efficient potential allergen detecting systems.

Acknowledgment:
We are grateful to Dr. V. K. Jayaraman for providing help on the application of data mining techniques. We acknowledge Rajeev Gangal, Director, Insilico Consulting, Pune, India for providing various resources to accomplish this research work. We thank G. Pugalenthi, Dilip Narayanan and Aftab Ahmad Khan for their support and comments on the manuscript.

References:
[1] L. Taylor, & L. Hefle, *J. Allergy Clin. Immunol.*, 107: 765 (2001) [PMID: 11344340]
[2] M. B. Stadler & B. M. Stadler, *FASEB. J.*, 17: 1141 (2003) [PMID: 12709401]
[3] D. Soeria-Atmadja, *et al.*, *Int. Arch. Allergy Immunol.*, 133: 101 (2004) [PMID: 14739578]
[4] S. Saha, *et al.*, *Nucleic Acids Res.*, 34: W202 (2006)[PMID: 16844994]
[5] T. Riaz, *et al.*, *Bioinformatics*, 21: 2570 (2005) [PMID: 15746289]
[6] K. B. Li, *et al.*, *Bioinformatics*, 20: 2572 (2004) [PMID: 15117757]
[7] O. Ivanciuc, *Nucleic Acids Res.*, 31: 359 (2003) [PMID: 12520022]
[8] D. T. Jones, *J Mol Biol.*, 292: 195 (1999) [PMID: 10493868]
[9] K. C. Chou & H. B. Shen, *Biochem Biophys Res Comm*., 360: 339 (2007) [PMID: 17586467]
[10] J. Salomon & D. R. Flower, *BMC Bioinformatics*, 7: 501 (2006) [PMID: 17105666]
[11] S. F. Altschul , *et al.*, *Nucleic Acids Res.*, 25: 3389 (1997) [PMID: 9254694]
[12] V. Vapnik, *The nature of statistical learning theory*, *Springer, New York* (1995)
[13] K. C. Chou, *Protein: Struc. Func. and Genetics*, 43: 246 (2001) [PMID: 11288174]
[14] D. Xie, *et al.*, *Nucleic Acids Res.*, 33: W105 (2005) [PMID: 15980436]
[15] K. C. Chou & Y. D. Cai, *J. Biol. Chem.*, 277: 45765 (2002) [PMID: 12186861]
[16] http://fermi.utmb.edu/SDAP
[17] http://www.csie.ntu.edu.tw/~cjlin/libsvm
[18] http://www.chou.med.harvard.edu/bioinf/PseAA/
[19] http://www.ncbi.nlm.nih.gov/BLAST/download.shtml

Edited by P. Kangueane
Citation: Kumar & Shelokar, Bioinformation 2(6): 253-256 (2008)