[Frontiers in Bioscience 13, 6072-6078, May 1, 2008]

Prediction of protein allergenicity using local description of amino acid sequence

Joo Chuan Tong[1], Martti T. Tammi[2,3,4]

[1]Data Mining Department, Institute for Infocomm Research, 21 Heng Mui Keng Terrace, Singapore 119613, [2]Department of Biological Sciences and Department of Biochemistry, National University of Singapore, 14 Science Drive 4, Singapore 117543, [3]Genome Institute, National Center for Genetic Engineering and Biotechnology, 113 Thailand Science Park, Paholyothin Rd., Klong 1, Klong Luang, Pathumthani, 12120 Thailand.[4]Karolinska Institutet, Department of Microbiology, Tumor and Cell Biology, Stockholm, Sweden

TABLE OF CONTENTS

1. ABSTRACT

The constant increase in atopic allergy and other hypersensitivity reactions has intensified The constant increase in atopic allergy and other hypersensitivity reactions has intensified the need for successful therapeutic approaches. Existing bioinformatic tools for predicting allergenic potential are primarily based on sequence similarity searches along the entire protein sequence and do not address the dual issues of conformational and overlapping B-cell epitope recognition sites. In this study, we report AllerPred, a computational system that is capable of capturing multiple overlapping continuous and discontinuous B-cell epitope binding patterns in allergenic proteins using SVM as its prediction engine. A novel representation of local protein sequence descriptors enables the system to model multiple overlapping continuous and discontinuous B-cell epitope binding patterns within a protein sequence. The model was rigorously trained and tested using 669 IUIS allergens and 1237 non-allergens. Testing results showed that the area under the receiver operating curve (A_{ROC}) of SVM models is 0.81 with 76% sensitivity at specificity of 76%. This approach consistently outperforms existing allergenicity prediction systems using a standardized testing dataset of experimentally validated allergens and non-allergen sequences.

2. INTRODUCTION

Atopic allergy and other hypersensitivity reactions are a major cause of chronic ill health in effluent industrial nations, affecting up to 25% of the general population (1-3). Allergy is caused by adverse immunologic reaction to causative agents known as allergens that are otherwise innocuous in nature. The acute symptoms of allergy are usually due to the release of inflammatory mediators when an allergen cross-links immunoglobulin E (IgE) antibodies on mast cells or basophils (4). This may be followed by a late-phase reaction characterized by the influx of T-cells, eosinophils and monocytes (5). Atopic individuals may have one or more manifestations of the disease including asthma, conjunctivitis, dermatitis (eczema), rhinitis (hay fever) and the severe anaphylaxis.

Assessment of potential allergenicity is an essential issue whenever new proteins are brought into contact with humans, either through food, or other modes of exposure. The current joint recommendation by the World Health Organization (WHO) and Food and Agriculture Organization (FAO) is a scheme based on a decision tree which compares local sequence similarity of a query protein against known allergenic proteins (6). Two

Table 1. Groupings of amino acid residues according to assigned features

Properties	Range	Amino Acids	References
Charge	1.00 0.00 -1.00	KRH ED STAGPQNMCLIVFWY	35
Hydrophobicity	1.80 - 4.50 -1.60 - 0.40 -4.00 - 3.20	CVLIMFA GSTWYP HQNEDKR	36
Polarity	0.00 – 0.35 1.43 – 3.53 49.50 – 52.00	AGILFV CMPSTYNQW RDHEK	37
Polarizability	0.00 – 0.14 0.15 – 0.23 0.22 – 0.41	GASNDCPTV EQILHKM FRYW	38
Bulkiness	3.40 – 9.47 11.50 – 13.69 14.28 – 21.67	GS ADNCEH RQKTMPYFILVW	37
Relative mutability	18.00 – 20.00 40.00 – 74.00 93.00 – 134.00	WC LFYGKPRHV QMITAEDSN	39
Solvent accessibility	0.32 – 0.51 0.66 – 0.71 0.78 – 0.93	WAMFLVIC TSYHG PDEQNKR	40
Normalized van der Waals volume	0.00 – 2.43 2.78 – 4.43 4.43 – 8.08	GASC PTDNVEQILM HKFRYW	35

decision criteria have been proposed for assessment of allergenic potential: identity of six or more contiguous amino acids, or minimum 35% sequence similarity over a window of 80 amino acids. Numerous research groups, including Fiers *et al*., Gendel *et al*., and Stadler and Stadler, developed computational tools that scan sequences that satisfy these criteria (7-10). While useful in some cases, the precision is low for methods solely relying on the six amino acid rule (11, 12).

More sophisticated bioinformatic tools for detecting motifs among allergenic sequences have been recently described. Zorzet and coworkers combined FASTA3 algorithm with k-Nearest-Neighbour (kNN) classifier to assess potential food protein allergenicity (13). Soeria-Atmadja and colleagues extended the study on a larger set of allergens using a combination of kNN classifier, Bayesian linear Gaussian classifier and Bayesian quadratic Gaussian classifier (2). Li *et al*. demonstrated the use of wavelet transform to predict potential allergens (14). Björklund *et al*. introduced the use of allergen-representative peptides (ARPs) for detection of potentially allergenic proteins (15). Saha and Raghava developed hybrid techniques combining probability matrices, IgE sequence comparison, ARPs and SVM for screening allergenic proteins (16). Although these are excellent attempts for assessing the potential allergenicity of protein sequences, an effective model to address the dual issues of conformational and overlapping B-cell epitope recognition sites is still currently lacking (17, 18).

In the present study, we report AllerPred, a prediction system for assessment of potential allergenicity of protein sequences. A novel data encoding scheme enables AllerPred to model multiple overlapping continuous and discontinuous B-cell epitope binding patterns within a protein sequence. The system is trained using official allergens approved by the International Union of Immunological Societies (IUIS) Allergen Nomenclature Sub-Committee plus non-allergens commonly found in consumed food with no records in existing allergen databases, and tested on experimentally validated allergens and non-allergen sequences. The effectiveness of the new encoding scheme in the representation of B-cell immunogenic regions is evaluated. The performance of AllerPred in the prediction of allergens from distantly related protein families is also assessed.

3. SYSTEM AND METHODS

3.1. Data

The dataset comprises 1906 (669 allergens and 1237 non-allergens) sequences. The available dataset is divided into training and testing datasets.

The training dataset consists of 631 IUIS approved allergens from the ALLERDB database (http://research.i2r.a-star.edu.sg/ Templar/DB/Allergen/) and 1219 non-allergens derived from http://www.slv.se/templates/SLV_Page.aspx?id=9343 using a debiasing strategy based on sequence similarity of protein sequences commonly found in consumed food with no records in existing allergen databases (16). The percentage of allergens represents ~34% of the testing dataset, while non-allergens represent the remaining 66%.

The testing dataset includes 38 IUIS allergens and 18 experimentally validated non-allergens extracted from the literature (19-29).

3.2. Model

The support vector machine (SVM) algorithm was used as implemented in SVMlight software (30). A comprehensive coverage of SVM has been covered in the literature (31, 32). In brief, SVMs belong to a class of statistical learning methods based on the structural risk minimization principle. The inputs to the SVM are binary strings or feature vectors representing encoded

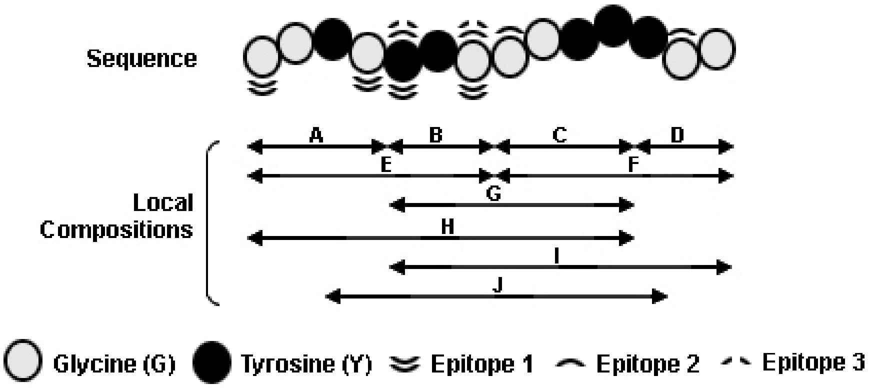

Figure 1. Schematic diagram of example local patterns currently modeled by AllerPred. Each protein sequence is divided into 10 subsequences (regions A – J) of varying length or composition to represent multiple overlapping continuous and discontinuous epitopes. For example, epitope 1 is represented by region E, epitope 2 is encapsulated by region I, and epitope 3 is effectively represented by region B. The feature vector for SVM training is derived by combining all descriptors from all local regions.

representations of amino acid attributes previously reported as significant for characterization of protein families. These include charge, hydrophobicity, polarity, polarizability, bulkiness, relative mutability, solvent accessibility and normalized van der Waals volume (Table 1). For each property, amino acids were divided into three classes according to their physico-chemical properties (33, 34). Parameters are trained by mapping input vectors into a high dimensional feature space and constructing an optimal separating hyperplane in the new feature space. The optimal separating hyperplane maximizes the margin between the positive and negative datasets and uniquely classifies the data into positive and negative examples. The trade-off parameter was set to 1000, to allow for imperfect separators with margins (30). Different kernel functions (linear, polynomial, radial, and sigmoid) were explored to optimize the prediction accuracy of the SVM models.

A novel representation of local protein descriptors is used to describe the physico-chemical properties of proteins. Each protein sequence is divided into 10 local regions of varying length and composition to describe both continuous and discontinuous epitopes (Figure 1). Three descriptors, composition (*C*), transition (*T*) and distribution (*D*), are used to represent the characteristics of each local region (Figure 2). *C* represents the composition of a given amino acid property by measuring the percentage of residues containing a particular property along a specified region. *T* stands for the percentage frequency with which a particular property changes along the entire region. *D* characterizes the distribution pattern of a particular property along the entire region by measuring the location of the first, 25, 50, 75 and 100% of residues with the property (32, 34). The descriptors for all local regions were combined to represent the general characteristics of the protein sequence and used as a feature vector for input into SVM.

For example, consider a hypothetical protein sequence “GGYGYYGGGYYYGG” containing 8 glycines and 6 tyrosines (Figure 1). Let region E be a subsequence denoting “GGYGYYG”. Let n_1 be the number of small amino acid residues and n_2 be the number of large amino acid residues within a specific region. The compositions for small residues (G; $n_1 = 4$) and large residues (Y; $n_2 = 3$) in region E are $n_1 / (n_1 + n_2) \times 100.00 = 57.14$ and $n_2 / (n_1 + n_2) \times 100.00 = 42.85$, respectively. The compositions in the other regions can be calculated in a similar manner. The *T* descriptor measures the percent frequency with which there is a transition from small to large residues or from large to small residues in each region. In region E, there are 4 transitions between small and large residues with a percent frequency $(4/6) \times 100.00 = 66.67$. The transitions for all other regions can be calculated in the same way. The 1st, 25, 50, 75 and 100% of small residues within region E are located within the first 1, 1, 2, 4 and 7 residues, respectively. The *D* descriptor for small residues is thus $1/7 \times 100.00 = 14.29$, $1/7 \text{ x } 100.00 = 14.29$, $2/7 \times 100.00 = 28.57$, $4/7 \times 100.00 = 57.14$, $7/7 \times 100.00 = 100.00$. The corresponding *D* descriptor for large residues can be calculated similarly. All three descriptors (*C*, *T* and *D*) from all local regions (A – J) were calculated, combined, and used as feature vector for SVM training. In AllerPred, amino acids were divided into three classes for each property, and a total of 21 descriptors are used to describe each attribute: 3 for C, 3 for T and 15 (3×5) for D (33).

3.3. Model evaluation

For each kernel function, 10-fold internal cross-validation was performed to assess to quality of the model (41). In *k*-fold cross-validation, *k* random, (approximately) equal-sized, disjoint partitions of the sample data are constructed, and a given model is trained on (*k*-1) partitions and tested on the excluded partition. The results are averaged after *k* such experiments, and the observed error rate may be taken as an estimate of the error rate expected upon generalization to new data.

The predictive performance of each model was assessed using sensitivity (SE), specificity (SP) and receiver operating characteristic (ROC) analysis as described previously (41). SE=TP/(TP+FN) and

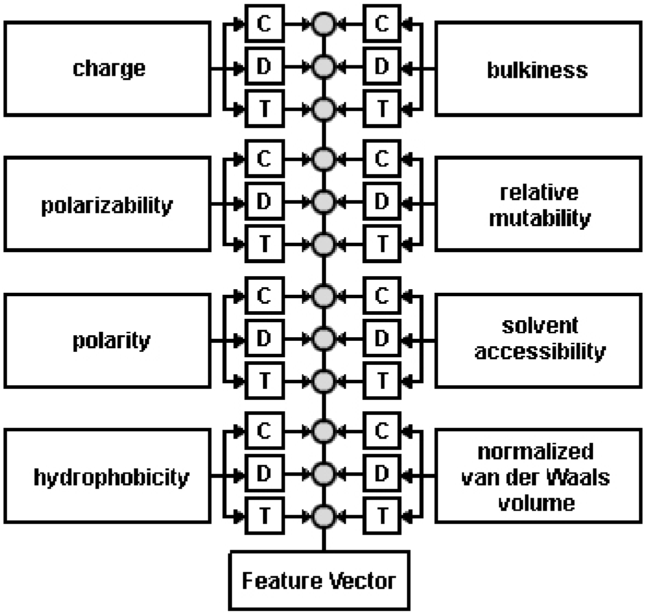

Figure 2. General scheme for representation of amino acid attributes in each local region. Three descriptors (*C*, *D*, *T*; squares) are used to represent each amino acid property (rectangles). All descriptors are combined and used to represent the feature vector of a local region.

SP=TN/(TN+FP), represent percentages of correctly predicted allergens and non-allergens, respectively. TP (true positives) stands for allergens correctly predicted as allergens and TN (true negatives) for non-allergens correctly predicted as non-allergens. FN (false negatives) refers to allergens predicted as non-allergens and FP (false positives) represents non-allergens predicted as allergens. The accuracy of our predictions was assessed by ROC analysis where the ROC curve is generated by plotting SE as a function of (1-SP) for various classification thresholds. The area under the ROC curve (A_{ROC}) provides a measure of overall prediction accuracy, A_{ROC}<70% for poor, A_{ROC}>80% for good and A_{ROC}>90% for excellent predictions.

4. RESULTS

The predictive performances of different kernel functions (linear, polynomial, radial, and sigmoid) were compared. AllerPred is based on the third degree polynomial kernel function encoded using descriptors derived from amino acid composition. The A_{ROC} value is 0.81. Using amino acid composition as input for training and testing, the system can predict allergenic proteins with SE of 76.00% and SP of 76.00%.

Although several allergenicity prediction systems have been described in the literature, only a limited number are available to the public (16). To benchmark our system, a standardized testing dataset comprising of 38 IUIS allergens and 18 experimentally validated non-allergens were used to evaluate three available techniques – wavelet transform models (14), SVM models based on global descriptors (33, 34) and sequence similarity search based on FAO/WHO Codex alimentarius guidelines (7). Our results indicate that, AllerPred, which utilizes local sequence descriptors for training SVM models (A_{ROC}=0.81), consistently outperforms SVM models based on global sequence descriptors (A_{ROC}=0.71), wavelet analysis (A_{ROC}=0.69) and FAO/WHO sequence similarity search (A_{ROC}=0.58) (7).

Collectively, our experiments indicate that we have developed a model that can make accurate predictions for potential allergenicity of proteins that has been validated using IUIS allergens and experimentally validated non-allergen sequences.

5. DISCUSSION

It has been well established that ~90% of B-cell epitopes are conformational in nature, where distant residues are brought into spatial proximity by protein folding (17, 42). In addition, the presence of overlapping epitopes on the surface of antigens has also been reported (18). However, this phenomenon has not been taken into account in current allergenicity prediction systems. Existing sequence similarity searching approaches have average predictivity (AROC≈0.70) when tested on experimentally validated allergens and non-allergens (33, 34). Low predictivity (AROC=0.58) is also observed for the FAO approach, consistent with previous assessment by other groups (10). It has also been reported that simplified profiles based on standard amino acid physico-chemical properties are not always able to describe with enough precision the protein sequence that needs to be modeled (43). This indicates that existing methodologies may not be effective in identifying B-cell epitope binding patterns. For practical applications of computational modeling techniques, it is essential to model every aspect of the immune responses (44).

Our strategy for prediction of protein allergenicity is to identify both continuous and discontinuous (possibly overlapping) immunodominant regions within an allergen. As illustrated in Figure 1, numerous epitopes may overlap along a protein sequence. Such phenomenon cannot be effectively modeled by existing computational strategies and sequence-based encoding schemes which focus on non-overlapping sequence similarity searches. Given the conformational nature of B-cell epitopes, the proposed scheme amplifies regional weights through the use of a series of local descriptors instead of a universal weighting scheme employed by existing techniques. In the current study, 10 local descriptors of varying length and composition were selected to capture the local characteristics of allergenic proteins. The SVM-based predictive technique based on this encoding scheme consistently outperforms current state-of-the-art. Future work will focus on optimizing the distribution of local descriptors in accordance with the experimentally validated binding patterns of B-cell epitopes. Given the complex nature of B-cell epitopes, the methodology proposed herein may be a possible solution towards effective modeling of continuous and discontinuous B-cell immunological regions.

6. ACKNOWLEDGEMENTS

JCT and MTT contributed equally to this article. The authors thank Dr Kuo-Bin Li for providing wavelet transform software to predict allergenic potential of proteins.

7. REFERENCES

1. Mekori Y. A.: Introduction to allergic diseases. *Crit Rev Food Sci Nutr* 36, S1-18 (1996)

2. Soeria-Atmadja D., A. Zorzet, M. G. Gustafsson, U. Hammerling: Statistical evaluation of local alignment features predicting allergenicity using supervised classification algorithms. *Int Arch Allergy Immunol* 133, 101-112 (2004)

3. Nieuwenhuizen N. E., A. L. Lopata: Fighting food allergy: current approaches. *Ann N Y Acad Sci* 1056, 30-45 (2005)

4. Sutton B. J., H. J. Gould: The human IgE network. *Nature* 366, 421-428 (1993)

5. Gould H. J., B. J. Sutton, A. J. Beavil, R. L. Beavil, N. McCloskey, H. A. Coker, D. Fear, L. Smurthwaite: The biology of IGE and the basis of allergic disease. *Annu Rev Immunol* 21, 579-628 (2003)

6. FAO/WHO: Codex Principles and Guidelines on Foods Derived from Biotechnology (2003)

7. Fiers M. W., G. A. Kleter, H. Nijland, A. A. Peijnenburg, J. P. Nap, R. C. van Ham: Allermatch, a webtool for the prediction of potential allergenicity according to current FAO/WHO Codex alimentarius guidelines. *BMC Bioinformatics* 5, 133 (2004)

8. Gendel S. M.: The use of amino acid sequence alignments to assess potential allergenicity of proteins used in genetically modified foods. *Adv Food Nutr Res* 42, 45-62 (1998)

9. Gendel S. M.: Sequence analysis for assessing potential allergenicity. *Ann N Y Acad Sci* 964, 87-98 (2002)

10. Stadler M. B., B. M. Stadler: Allergenicity prediction by protein sequence. *FASEB J* 17, 1141-1143 (2003)

11. Hileman R. E., A. Silvanovich, R. E. Goodman, E. A. Rice, G. Holleschak, J. D. Astwood, S. L. Hefle: Bioinformatic methods for allergenicity assessment using a comprehensive allergen database. *Int Arch Allergy Immunol* 128, 280-291 (2002)

12. Silvanovich A., M. A. Nemeth, P. Song, R. Herman, L. Tagliani, G. A. Bannon: The value of short amino acid sequence matches for prediction of protein allergenicity. *Toxicol Sci* 90, 252-258 (2006)

13. Zorzet A., M. Gustafsson, U. Hammerling: Prediction of food protein allergenicity: a bioinformatic learning systems approach. *In Silico Biol* 2, 525-534 (2002)

14. Li K. B., P. Issac, A. Krishnan: Predicting allergenic proteins using wavelet transform. *Bioinformatics* 20, 2572-2578 (2004)

15. Bjorklund A. K., D. Soeria-Atmadja, A. Zorzet, U. Hammerling, M. G. Gustafsson: Supervised identification of allergen-representative peptides for in silico detection of potentially allergenic proteins. *Bioinformatics* 21, 39-50 (2005)

16. Saha S., G. P. Raghava: AlgPred: prediction of allergenic proteins and mapping of IgE epitopes. *Nucleic Acids Res* 34, W202-W209 (2006)

17. Ho J., K. S. MacDonald, B. H. Barber: Construction of recombinant targeting immunogens incorporating an HIV-1 neutralizing epitope into sites of differing conformational constraint. *Vaccine* 20, 1169-1180 (2002)

18. Guo J., R. S. McIntosh, B. Czarnocka, A. P. Weetman, B. Rapoport, S. M. McLachlan: Relationship between autoantibody epitopic recognition and immunoglobulin gene usage. *Clin Exp Immunol* 111, 408-14 (1998)

19. Chakraborty S., N. Chakraborty, A. Datta: Increased nutritive value of transgenic potato by expressing a nonallergenic seed albumin gene from Amaranthus hypochondriacus. *Proc Natl Acad Sci USA* 97, 3724-9 (2000)

20. Laffer S., S. Hamdi, C. Lupinek, W. R. Sperr, P. Valent, P. Verdino, W. Keller, M. Grote, K. Hoffmann-Sommergruber, O. Scheiner, D. Kraft, M. Rideau, R. Valenta: Molecular characterization of recombinant T1, a non-allergenic periwinkle (Catharanthus roseus) protein, with sequence similarity to the Bet v 1 plant allergen family. *Biochem J* 373, 261-269 (2003)

21. Epton M. J., W. Smith, B. J. Hales, L. Hazell, P. J. Thompson, W. R. Thomas: Non-allergenic antigen in allergic sensitization: responses to the mite ferritin heavy chain antigen by allergic and non-allergic subjects. *Clin Exp Allergy* 32, 1341-1347 (2002)

22. Ortona E., P. Margutti, F. Delunardo, S. Vaccari, R. Rigano, E. Profumo, B. Buttari, A. Teggi, A. Siracusano: Molecular and immunological characterization of the C-terminal region of a new Echinococcus granulosus Heat Shock Protein 70. *Parasite Immunol* 25, 119-126 (2003)

23. Szakos E., G. Lakos, M. Aleksza, E. Gyimesi, G. Pall, B. Fodor, J. Hunyadi, E. Solyom, S. Sipka: Association between the occurrence of the anticardiolipin IgM and mite allergen-specific IgE antibodies in children with extrinsic type of atopic eczema/dermatitis syndrome. *Allergy* 59, 164-167 (2004)

24. Siler D. J., K. Cornish, R. G. Hamilton: Absence of cross-reactivity of IgE antibodies from subjects allergic to Hevea brasiliensis latex with a new source of natural rubber latex from guayule (Parthenium argentatum). *J Allergy Clin Immunol* 98, 895-902 (1996)

25. Dearman R. J., I. Kimber: Determination of protein allergenicity: studies in mice. *Toxicol Lett* 120, 181-186 (2001)

26. Dearman R. J., S. Stone, H. T. Caddick, D. A. Basketter, I. Kimber: Evaluation of protein allergenic potential in mice: dose-response analyses. *Clin Exp Allergy* 33, 1586-1594 (2003)

27. Banerjee B., V. P. Kurup, P. A. Greenberger, K. J. Kelly, J. N. Fink: C-terminal cysteine residues determine the IgE binding of Aspergillus fumigatus allergen Asp f 2. *J Immunol* 169, 5137-5144 (2002)

28. Takai T., T. Yuuki, Y. Okumura, A. Mori, H. Okudaira: Determination of the N- and C-terminal sequences required to bind human IgE of the major house dust mite allergen Der f 2 and epitope mapping for monoclonal antibodies. *Mol Immunol* 34, 255-261 (1997)

29. Mine Y., E. Sasaki, J. W. Zhang: Reduction of antigenicity and allergenicity of genetically modified egg white allergen, ovomucoid third domain. *Biochem Biophys Res Commun* 302, 133-137 (2003)

30. Joachims T.: Learning to classify text using support vector machines. *Kluwer Academic Publishers, Boston* (2002)

31. Vapnik V. N.: Statistical learning theory. *Wiley, New York* (1998)

32. Zhang Z. H., J. L. Koh, G. L. Zhang, K. H. Choo, M. T. Tammi, J. C. Tong: AllerTool: a web server for predicting allergenicity and allergic cross-reactivity in proteins. *Bioinformatics* 23, 504-506 (2007)

33. Cui J., L. Y. Han, H. Li, C. Y. Ung, Z. Q. Tang, C. J. Zheng, Z. W. Cao, Y. Z. Chen: Computer prediction of allergen proteins from sequence-derived protein structural and physicochemical properties. *Mol Immunol* 44, 514-520 (2007)

34. Dubchak I., I. Muchnik, S. R. Holbrook, S. H. Kim: Prediction of protein folding class using global description of amino acid sequence. *Proc Natl Acad Sci USA* 92, 8700-8704 (1995)

35. Fauchere J. L., M. Charton, L. B. Kier, A. Verloop, V. Pliska: Amino acid side chain parameters for correlation studies in biology and Pharmacology. *Int J Peptide Protein Res* 32, 269-278 (1988)

36. Kyte J., R. F. Doolittle: A simple method for displaying the hydropathic character of a protein. *J Mol Biol* 157, 105-132 (1982)

37. Zimmerman J. M., N. Eliezer, R. Simha: The characterization of amino acid sequences in proteins by statistical methods. *J Theor Biol* 21, 170-201 (1968)

38. Charton M., B. I. Charton: The structural dependence of amino acid hydrophobicity parameters. *J Theor Biol* 99, 629-644 (1982)

39. Dayhoff M. O., R.M. Schwartz, B.C. Orcutt: In *Atlas of Protein Sequence and Structure* 5, Suppl 3 (1978)

40. Bordo D., P. Argos: Suggestions for "safe" residue substitutions in site-directed mutagenesis. *J Mol Biol* 217: 721-729 (1991)

41. Tong J. C., G. L. Zhang, T. W. Tan, J. T. August, V. Brusic, S. Ranganathan: Prediction of HLA-DQ3.2beta ligands: evidence of multiple registers in class II binding peptides. *Bioinformatics* 22, 1232-1238 (2006)

42. Barlow D. J., M. S. Edwards, J. M. Thornton: Continuous and discontinuous protein antigenic determinants. *Nature* 322, 747-748 (1986)

43. Bacardit J., M. Stout, J. D. Hirst, K. Sastry, X. Llora, N. Krasnogor: Automated alphabet reduction method with evolutionary algorithms for protein structure prediction. In *Proceedings of the 9th annual conference on Genetic and evolutionary computation*, ACM Press, 346-353 (2007)

44. Doytchinova I. A., D. R. Flower: Towards the in silico identification of class II restricted T-cell epitopes: a partial least squares iterative self-consistent algorithm for affinity prediction. *Bioinformatics* 19, 2263-2270 (2003)

Abbreviations: A_{ROC}: Area under the receiver operating curve, WHO: World Health Organization, FAO: Food and Agriculture Organization, kNN: k-Nearest-Neighbor, ARPs: allergen-representative peptides, IUIS: Immunological Societies, SVM: support vector machine, SE: sensitivity, SP: specificity, TP: true positives, TN: true negatives, FN: false negatives, FP: false positives, ROC: receiver operating characteristic

Key Words: Bioinformatics, Allergens, Allergy Assessment, Support Vector Machine, Prediction, Review

Send correspondence to: Martti Tammi, Department of Biological Sciences and Department of Biochemistry, National University of Singapore, 14 Science Drive 4, Singapore 117543, Tel: 65-6516-4255, Fax: 65-6779-2486, E-mail: martti@nus.edu.sg

http://www.bioscience.org/current/vol13.htm