Prediction of protein allergenicity based on signal-processing bioinformatics approach

Charalambos Chrysostomou[1]* and Huseyin Seker[2]

Abstract— **Current bioinformatics tools accomplish high accuracies in classifying allergenic protein sequences with high homology and generally perform poorly with low homology protein sequences. Although some homologous regions explained Immunoglobulin E (IgE) cross-reactivity in groups of allergens, no universal molecular structure could be associated with allergenicity. In addition, studies have showed that cross-reactivity is not directly linked to the homology between protein sequences. Therefore, a new homology independent method needs to be developed to determine if a protein is an allergen or not. The aim of this study is therefore to differentiate sets of allergenic and non-allergenic proteins using a signal-processing based bioinformatics approach.**

In this paper, a new method was proposed for characterisation and classification of allergenic protein sequences. For this method hydrophobicity amino acid index was used to encode proteins to numerical sequences and Discrete Fourier Transform to extract features for each protein. Finally, a classifier was constructed based on Support Vector Machines. In order to demonstrate the applicability of the proposed method 857 allergen and 1000 non-allergen proteins were collected from UniProt online database. The results obtained from the proposed method yielded: MCC: 0.752 $\pm$ 0.007, Specificity: 0.912 $\pm$ 0.005, Sensitivity: 0.835 $\pm$ 0.008 and Total Accuracy: 87.65% $\pm$ 0.004.

I. Introduction

An allergy is a Type I hypersensitivity and caused when a person's immune system overreacts to substances from the environment [1]. This reaction results in an inflammatory response from mild to life-threading symptoms. In recent years, bioinformatics tools have been developed for analysis and classification of allergenic protein sequences. Such tools utilize allergen representative peptides (ARPs) [2], sequence similarity search [3], Support Vector Machines (SVM) [3], [4] and k-Nearest-Neighbor (kNN) classifiers [5]. These bioinformatics tools accomplish high accuracies in classifying allergenic protein sequences with high homology and generally perform poorly with low homology protein sequences.

Although some homologous regions explained Immunoglobulin E (IgE) cross-reactivity in groups of allergens [6], no universal molecular structure could be associated with allergenicity as reported in previous studies [6], [7], [8], [9]. In addition, studies have showed that cross-reactivity is not directly linked to the homology between protein sequences [10], [11], [12]. Therefore, a new homology independent method needs to be developed to determine if a protein is an allergen or not. The aim of this study is to differentiate sets of allergenic and non-allergenic proteins using a homology independent signal-processing based bioinformatics approach.

In this paper, a novel method is presented which uses Discrete Fourier Transform to extract information from protein sequences and Support Vector machines as a predictive tool for allergenic protein sequences. Additionally, the collection of allergenic and non-allergenic protein sequences from UniProt [13], will be discussed. The paper is organised as follows: Section II presents the methods and materials developed and used, while Section III presents the results obtained. Finally, concluding remarks with discussions are stated in Section IV.

II. Materials and Methods

A. Processing the Protein Sequences

By using an amino acid index [14] the protein sequences can be converted to a numerical sequence, in order for Discrete Fourier Transform (DFT) [15] to be applied. Before applying DFT to the protein sequences, zero-padding and windowing techniques commonly used in signal processing need to be considered. Previous work has showed that by applying zero-padding and windowing methods to the converted protein sequences can influence the features extracted from signal processing techniques [16].

By using the windowing technique, where a pre-calculated window is multiplied to the protein sequence numerical sequences, spectral leakage is reduced. For the analysis of allergen protein sequences the Hamming window [17] is used as showed in Equation 1.

$$w = 0.54 - 0.46cos\left(\frac{2\pi n}{N-1}\right) \quad n = 0 \leq n \leq N-1 \quad (1)$$

The next step in the analysis of protein sequences using DFT is the zero-padding method, where a number of zero elements are added to the end of each sequence to increase signal length. Zero-padding is an important step in the analysis, as the protein sequences used may not have equal lengths.

The final step in processing the protein sequences is to apply DFT. The Discrete Fourier Transform (DFT) can be

[1]Department of Genetics, University of Leicester, University Road Leicester, LE1 7RH, United Kingdom

[2]Bio-Health Informatics Research Group, Centre for Computational Intelligence, Faculty of Technology, De Montfort University, Leicester, LE1 9BH, UK

cc390@le.ac.uk, hseker@dmu.ac.uk

*Corresponding Author

978-1-4244-7929-0/14/$26.00 ©2014 IEEE

$$X(n) = \sum_{m=0}^{N-1} x(m) e^{-j(2\pi/N)nm} \quad n = 0, 1, ..., N-1 \quad (2)$$

where $X(n)$ are the DFT coefficients, N is the total number of points in the series and $x(m)$ is the mth member of the numerical series. Henceforth only the $(N/2)$ points of the series will be used as the DFT coefficients contains two mirror parts.
The output of DFT is a complex sequence and can be formulated as

$$X(n) = (R(n) + jI(n)), \quad n = 0, 1, ..., (N-1)/2 \quad (3)$$

where $R(n)$ and $I(n)$ are the Real and Imaginary parts of the sequence, respectively.
The absolute spectrum can be formulated as

$$S_{(n)} = X(n)X^*(n) = |X(n)|^2, \quad n = 0, 1, ..., (N-1)/2 \quad (4)$$

where $X(n)$ are the DFT coefficients of the series $x(n)$, $X*(n)$ are the complex conjugates and $S_{(n)}$ is the absolute spectrum.

The coefficients from the absolute spectrum for each protein can be used as a feature set to represent the characteristics of allergen and non-allergen protein sequences.

B. Classification of Allergenic Protein Sequences based on Support Vector Machine

A support vector machine (SVM) [18], is a non-probabilistic linear classifier [18], [19] used for data analysis and pattern recognition analysis. In the literature, SVM is used in the analysis and classification of protein sequences with very promising results. Some of the research areas where SVM is successfully applied are protein interactions prediction [20], protein secondary structure prediction [21], RNA-binding proteins from primary sequence prediction [22] and protein subcellular localization prediction [23].

For this analysis, LIBSVM python library [24] was used to construct the classifier. The radial basis function (RBF) kernel function was used, as RBF has shown to be the simplest to adapt and the most generally applicable [25]. Finally, grid search is applied to find the optimal values of the kernel C and γ parameters.

C. Evaluating the Performance of the Predictive Model

For this analysis, the K-fold (5-fold) cross-validation technique [26] is used to assess the performance of the allergen classifier. This technique usually is used to approximate how these predictive models will behave and perform in practice. Cross-validation is important for independently testing and validating different theories on existing data, where collecting additional data is impossible, costly or time consuming.

The performance of the allergen classifier was evaluated based on sensitivity (SE), specificity (SP) and total accuracy (TACC). They can be calculated by using the following equations, respectively

$$SE = \frac{TP}{TP + FN} \quad (5)$$

$$SP = \frac{TN}{TN + FP} \quad (6)$$

$$TACC = \frac{TP + TN}{TP + TN + FP + FN} * 100 \quad (7)$$

where true positives (TP) and true negatives (TN) represent the correctly identified allergen and non-allergen protein sequences, respectively. In addition, false negatives (FN) and false positives (FP) represent the misidentified allergen and non-allergen protein sequences, respectively.

Additionally, three different evaluation methods, Matthews correlation coefficient (MCC) [27], G-mean [28] and F-measure [29], were also used that have been shown to provide more reliable results when evaluating a classifier.
The MCC can be calculated by using Equation 8

$$MCC = \frac{(TP * TN) - (FN * FP)}{\sqrt{(TN + FN)(TP + FN)(TN + FP)(TP + FP)}} \quad (8)$$

The output of MCC is a value between [-1,1] where 0 indicates random classification, 1 indicates 100% correct classification and -1 indicates 100% misclassification.
G-mean, is showed in Equation 9

$$GMean = \sqrt{SE \ * \ SP} \quad (9)$$

Finally, F-measure, is given in Equation 10

$$F - Measure = \frac{2 * TP}{2 * TP + FN + FP} \quad (10)$$

The output of the F-measure is a value between [0,1] where 1 represents 100% correct classification and 0 represents 100% misclassification.

D. Amino Acid Index

As described in the previous sections an amino acid index needs to be used to convert the proteins sequences to numerical sequences. Each amino acid index represents a physical characteristic of the protein or amino acid they represent. From the liturature, hydrophobicity feature has

been shown to be related to allergenic proteins [6], [7], [8], [9]. In this paper for the classification of allergenic and non-allergenic protein sequences hydrophobicity amino acid index will be used, as described in Table I.

TABLE I
HYDROPHOBICITY AMINO ACID INDEX

Amino Acid	Value
A	0.44
R	-2.42
N	-1.32
D	-0.31
C	0.58
Q	-0.71
E	-0.34
G	0
H	-0.01
I	2.46
L	2.46
K	-2.45
M	1.1
F	2.54
P	1.29
S	-0.84
T	-0.41
W	2.56
Y	1.63
V	1.73

E. Allergenic Protein Databases

For this analysis, data were collected from UniProt [13] (http://www.uniprot.org). From UniProt, only the verified Allergens were considered. Table II lists the number of allergen and non-allergen proteins collected from the database. Table II shows the number of protein sequences, as well as the maximum, minimum, and average length of the protein sequences.

TABLE II
ALLERGEN AND NON-ALLERGEN ONLINE DATABASES USED IN THIS STUDY

	Allergen Proteins	**Non Allergen Proteins**
Number of Proteins	857	1000
Maximum Length	1558	5890
Minimum Length	5	78
Average Length	235.65	752.56

III. RESULTS AND DISCUSSIONS

In this paper, a study is performed in order to differentiate sets of allergenic and non-allergenic proteins using a signal-processing based bioinformatics approach. For this analysis, DFT coefficients were used to characterise protein sequences as well as hydrophobicity amino acid index in order to encode protein sequences to numerical sequences. Finally, support vector machines were utilised as a predictive tool. In order to test how the prediction of allergenic proteins based on the proposed method, the following analysis was carried out.

From UniProt database 817 allergen and 1000 non-allergen proteins were used. The performance of the allergen classifier was evaluated based on MCC, sensitivity, specificity and total accuracy of the classifier. In order to ensure that the results are generalised, 5-fold cross-validation was used to train the classifier. This process was trained for 10 times and the average values along with standard deviation were presented.

All the analyses carried out resulted in an average predictive accuracy of 87.65% $\pm$ 0.004. In addition, MCC, specificity and sensitivity values are found to be 0.752 $\pm$ 0.007, 0.912 $\pm$ 0.005 and 0.835 $\pm$ 0.008, respectively. All the results are given in Table III.

The results obtained using the proposed method indicate that allergens and non-allergen protein sequences can be accurately classified. As current bioinformatics tools use homology to classify allergenic protein sequences, they accomplish high accuracies in classifying sequences with high homology and generally perform poorly with low homology sequences. From the literature no indications exist that cross-reactivity and homology of protein sequences are linked [10], [11], [12], and no universal molecular structure of allergen protein sequences currently exist [6], [7], [8], [9]. The advantages of the proposed method over the existing tolls are that the method developed in this paper is a non-parametric and homology independent method that can be directly linked to physical characteristics of the protein sequence, as as an example that hydrophobicity amino acid index was used to encode the sequences.

IV. CONCLUSIONS

In this paper, a new method was proposed for characterisation and classification of allergenic protein sequences. For this method hydrophobicity amino acid index was used to encode proteins to numerical sequences and Discrete Fourier Transform to extract features for each protein. Finally, a classifier was constructed based on Support Vector Machines. In order to demonstrate the applicability of the proposed method 857 allergen and 1000 non-allergen proteins were collected from UniProt online database.

In the literature more that 500 amino acid indices exist [14], which can be used to encode protein sequences to numerical sequences. Further research needs to be carried out with different amino acid indices, and compare the results with those of the hydrophobicity. Furthermore, larger datasets should be collected from various allergenic databases that exist, and the accuracy of the method needs to be tested on different homology datasets in order to be able to demonstrate further applicability and generalizability of the methods described in the paper.

REFERENCES

[1] A. B. Kay, "Overview of allergy and allergic diseases: with a view to the future," *British medical bulletin*, vol. 56, no. 4, pp. 843–864, 2000.

[2] Å. Björklund, D. Soeria-Atmadja, A. Zorzet, U. Hammerling, and M. Gustafsson, "Supervised identification of allergen-representative peptides for in silico detection of potentially allergenic proteins," *Bioinformatics*, vol. 21, no. 1, p. 39, 2005.

TABLE III
RESULTS OF THE ANALYSIS

	Specificity	Sensitivity	Total Accuracy	MCC	G-Mean	F-Measure
Analysis 1	0.912	0.847	88.21%	0.763	0.879	0.869
Analysis 2	0.917	0.831	87.72%	0.753	0.873	0.862
Analysis 3	0.911	0.823	87.02%	0.739	0.866	0.854
Analysis 4	0.915	0.841	88.10%	0.761	0.877	0.867
Analysis 5	0.905	0.839	87.45%	0.747	0.871	0.861
Analysis 6	0.916	0.828	87.56%	0.75	0.871	0.86
Analysis 7	0.913	0.827	87.35%	0.746	0.869	0.858
Analysis 8	0.912	0.838	87.78%	0.754	0.874	0.863
Analysis 9	0.901	0.841	87.35%	0.745	0.871	0.86
Analysis 10	0.915	0.838	87.94%	0.758	0.876	0.865
Average Results	0.9117 ± 0.005	0.8353 ± 0.008	$87.65\% \pm 0.004$	0.7516 ± 0.008	0.8727 ± 0.004	0.8619 ± 0.004

[3] H. Muh, J. Tong, and M. Tammi, "Allerhunter: a svm-pairwise system for assessment of allergenicity and allergic cross-reactivity in proteins," *PloS one*, vol. 4, no. 6, p. e5861, 2009.

[4] J. Cui, L. Han, H. Li, C. Ung, Z. Tang, C. Zheng, Z. Cao, and Y. Chen, "Computer prediction of allergen proteins from sequence-derived protein structural and physicochemical properties," *Molecular immunology*, vol. 44, no. 4, pp. 514–520, 2007.

[5] A. Zorzet, M. Gustafsson, and U. Hammerling, "Prediction of food protein allergenicity: A bio-informatic learning systems approach," *In Silico Biology*, vol. 2, no. 4, pp. 525–534, 2002.

[6] J. Wal, "Structure and function of milk allergens," *Allergy*, vol. 56, pp. 35–38, 2001.

[7] R. Furmonaviciene, B. Sutton, C. Laughton, H. Sewell, and F. Shakib, "The definition of allergen-specific molecular surface features: new insights into allergenicity," *Bioinformatics*, vol. 21, pp. 4201–4204, 2005.

[8] K. Mengumpun, C. Tayapiwatana, R. Hamilton, P. Sangsupawanich, and R. Wititsuwannakul, "Hydrophobic allergens from the bottom fraction membrane of hevea brasiliensis," *Asian Pacific Journal of Allergy and Immunology*, vol. 26, no. 2-3, pp. 129–136, 2010.

[9] M. Gijzen, S. Miller, K. Kuflu, R. Buzzell, and B. Miki, "Hydrophobic protein synthesized in the pod endocarp adheres to the seed surface," *Plant physiology*, vol. 120, no. 4, p. 951, 1999.

[10] R. Aalberse, "Structural biology of allergens," *Journal of allergy and clinical immunology*, vol. 106, no. 2, pp. 228–238, 2000.

[11] W. Thomas, B. Hales, and W. Smith, "Structural biology of allergens," *Current Allergy and Asthma Reports*, vol. 5, no. 5, pp. 388–393, 2005.

[12] M. Chapman, A. Pomés, H. Breiteneder, and F. Ferreira, "Nomenclature and structural biology of allergens," *Journal of allergy and clinical immunology*, vol. 119, no. 2, pp. 414–420, 2007.

[13] A. Bairoch, R. Apweiler, C. Wu, W. Barker, B. Boeckmann, S. Ferro, E. Gasteiger, H. Huang, R. Lopez, M. Magrane, *et al.*, "The universal protein resource (uniprot)," *Nucleic acids research*, vol. 33, no. suppl 1, pp. D154–D159, 2005.

[14] S. Kawashima, P. Pokarowski, M. Pokarowska, A. Kolinski, T. Katayama, and M. Kanehisa, "Aaindex: amino acid index database, progress report 2008," *Nucleic acids research*, vol. 36, no. suppl 1, p. D202, 2008.

[15] D. Sundararajan, *The discrete Fourier transform: theory, algorithms and applications*. World Scientific Pub Co Inc, 2001.

[16] C. Chrysostomou, H. Seker, and N. Aydin, "Effects of windowing and zero-padding on complex resonant recognition model for protein sequence analysis," in *Proceedings of EMBC 2011*, Boston, USA, August 2011, pp. 4955–8.

[17] R. Blackman and J. Tukey, "The Measurement of Power Spectra, 190 pp," *New York*, 1958.

[18] B. E. Boser and et al., "A training algorithm for optimal margin classifiers," in *PROCEEDINGS OF THE 5TH ANNUAL ACM WORKSHOP ON COMPUTATIONAL LEARNING THEORY*. ACM Press, 1992, pp. 144–152.

[19] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[20] S. Lo, C. Cai, Y. Chen, and M. Chung, "Effect of training datasets on support vector machine prediction of protein-protein interactions," *Proteomics*, vol. 5, no. 4, pp. 876–884, 2005.

[21] C. Chen, L. Chen, X. Zou, and P. Cai, "Prediction of protein secondary structure content by using the concept of chous pseudo amino acid composition and support vector machine," *Protein and peptide letters*, vol. 16, no. 1, pp. 27–31, 2009.

[22] L. Han, C. CaiI, S. Lo, M. Chung, and Y. Chen, "Prediction of RNA-binding proteins from primary sequence by a support vector machine approach," *Rna*, vol. 10, no. 3, p. 355, 2004.

[23] S. Hua and Z. Sun, "Support vector machine approach for protein subcellular localization prediction," *Bioinformatics*, vol. 17, no. 8, pp. 721–728, 2001.

[24] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 27:1–27:27, 2011, software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm.

[25] R. Christensen, M. Enuameh, M. Noyes, M. Brodsky, S. Wolfe, and G. Stormo, "Recognition models to predict dna-binding specificities of homeodomain proteins," *Bioinformatics*, vol. 28, no. 12, pp. i84–i89, 2012.

[26] I. Witten, E. Frank, and M. Hall, *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2011.

[27] B. Matthews *et al.*, "Comparison of the predicted and observed secondary structure of t4 phage lysozyme." *Biochimica et biophysica acta*, vol. 405, no. 2, p. 442, 1975.

[28] M. Abramowitz and I. Stegun, *Handbook of mathematical functions with formulas, graphs, and mathematical tables*. Dover publications, 1964, vol. 55, no. 1972.

[29] C. J. van Rijsbergen, *Information Retrieval*. Butterworth, 1979.