

Computer prediction of allergen proteins from sequence-derived protein structural and physicochemical properties

Juan Cui^a, Lian Yi Han^a, Hu Li^a, Choong Yong Ung^a,
Zhi Qun Tang^a, Chan Juan Zheng^a, Zhi Wei Cao^b, Yu Zong Chen^{a,b,*}

^a *Bioinformatics and Drug Design Group, Department of Pharmacy and Computational Science, National University of Singapore, Blk SoC 1, Level 7, 3 Science Drive 2, Singapore 117543, Singapore*

^b *Shanghai Center for Bioinformatics Technology, 100, Qinzhou Road, Shanghai 200235, PR China*

Received 7 November 2005; received in revised form 6 February 2006; accepted 14 February 2006

Available online 23 March 2006

Abstract

Background: Computational methods have been developed for predicting allergen proteins from sequence segments that show identity, homology, or motif match to a known allergen. These methods achieve good prediction accuracies, but are less effective for novel proteins with no similarity to any known allergen.

Methods: This work tests the feasibility of using a statistical learning method, support vector machines, as such a method. The prediction system is trained and tested by using 1005 allergen proteins from the Allergome database and 22,469 non-allergen proteins from 7871 Pfam families.

Results: Testing results by an independent set of 229 allergen and 6717 non-allergen proteins from 7871 Pfam families show that 93.0% and 99.9% of these are correctly predicted, which are comparable to the best results of other methods. Of the 18 novel allergen proteins non-homologous to any other proteins in the Swissprot database, 88.9% is correctly predicted. A further screening of 168,128 proteins in the Swissprot database finds that 2.9% of the proteins are predicted as allergen proteins, which is consistent with the estimated numbers from motif-based methods.

Conclusions: Our study suggests that SVM is a potentially useful method for predicting allergen proteins and it has certain capability for predicting novel allergen proteins. Our software can be accessed at <http://jing.cz3.nus.edu.sg/cgi-bin/APPEL>.

© 2006 Elsevier Ltd. All rights reserved.

Keywords: Allergen; Immunology; Statistical learning method; Support vector machine

1. Introduction

Allergic diseases are responsible for such serious health problems as asthma, dermatitis, food allergy, bee sting allergy, conjunctivitis, and severe systemic anaphylaxis (Gould et al., 2003; Ono, 2000; Sampson, 2004). These diseases affect a substantial portion of the general population. Asthma and atopic dermatitis, respectively, affect 10% and 15% of the children in some countries (Cookson, 2004). Fish allergy and general food allergy were reported in 2.3% (Sicherer et al., 2004) and 4% (Sampson, 2004) of the US population, respectively. Overall, up to one third of the human population are affected by one or more of these diseases (Ono, 2000). A considerable portion of the health care expenses has been directed at the treatment of allergic diseases. Consider-

able efforts have been made for identifying factors contributing to these diseases (Bochner and Busse, 2004; Breiteneder and Mills, 2005b; Lack, 2004; Ono, 2000).

An allergic reaction is initiated when an allergen interacts with an IgE antibody (Sutton and Gould, 1993). Many of the allergens are proteins and the prediction of these allergen proteins facilitate the study of molecular mechanisms and risk assessment of allergic reactions in different individuals (Breiteneder and Mills, 2005b; Malandain, 2004). A consensus view since the 1970s and 1980s is that any non-self molecule with antigenic properties could potentially be allergenic, and the risk of developing allergic sensitization is mainly determined by the host's immune responsiveness (Cebra et al., 1984; Stewart, 1973), genetic and developmental factors (Bochner and Busse, 2004; Gould et al., 2003), route and frequency of exposure (Mills et al., 2004), and in some cases simultaneous exposure to co-factors (van Wijk et al., 2005; Venge et al., 1996). Non-the-less, groups of allergen proteins have various levels of

* Corresponding author. Tel.: +65 6516 6877; fax: +65 6774 6756.
E-mail address: yzchen@cz3.nus.edu.sg (Y.Z. Chen).

common structural and physicochemical features (Bredehorst and David, 2001; Breiteneder and Clare Mills, 2005; Breiteneder and Ebner, 2000; Breiteneder and Ebner, 2001; Breiteneder and Mills, 2005a,b; Davis et al., 2001; Huby et al., 2000; Wal, 2001) that have been explored for predicting allergenicity from protein sequence (Brusic et al., 2003; Hileman et al., 2002; Ivanciuc et al., 2003; Kleter and Peijnenburg, 2002; Soeria-Atmadja et al., 2004; Stadler and Stadler, 2003).

Allergen proteins tend to be ovoid in shape (Huby et al., 2000) with a molecular mass of 5000–70,000 (Bredehorst and David, 2001). They have been frequently found to contain IgE epitopes or specific structural integrity for IgE binding (Bredehorst and David, 2001; Wal, 2001). They are expected to possess certain structural and biological attributes that preserve their structure from the destructive effects of low pH, proteolysis, surfactants and thermal processing (Davis et al., 2001; Mills et al., 2004). The structure of some allergens may be stabilized by disulfide bonds, glycosylation, glycation, structural aggregation, and appearance of repetitive structures (Breiteneder and Mills, 2005b). Many allergens bind to lipids and membranes to enable transportation and protection from destructive environment (Breiteneder and Mills, 2005a,b; Mills et al., 2004). Many known plant allergens belong to a few protein families (Breiteneder and Clare Mills, 2005; Breiteneder and Ebner, 2001) and are homologous to pathogenesis-related proteins (Breiteneder and Ebner, 2000). While individual features are not the sole determinant (Mills et al., 2004), some collective combination of these features likely play important roles in allergenicity. Analysis of the sequence of allergen proteins that

encodes these features may therefore provide useful clues to the prediction of their allergenic potential.

Several computational methods have been introduced as tools for predicting allergenic potentials of proteins from their sequence (Brusic et al., 2003). One approach, the FAO/WHO method, is based on the search of protein sequences that have sequence identity over a segment of at least 6–8 amino acids or have 35% similarity over an 80 amino acid window to a known allergen protein (Hileman et al., 2002). Another strategy is to predict allergy proteins based on the 52 sequence motifs derived from sequence analysis of known allergen proteins (Stadler and Stadler, 2003). Methods that combine or integrate these and/or other information have been introduced (Ivanciuc et al., 2003; Kleter and Peijnenburg, 2002; Soeria-Atmadja et al., 2004). Statistical learning methods, such as kNN and linear/quadratic Gaussian, have also been explored for predicting allergen proteins from locally aligned sequences (Soeria-Atmadja et al., 2004).

Table 1 summarizes the computational methods explored and the reported prediction accuracies. These methods predict allergen proteins at impressive accuracies of 75–100%. The accuracies for two non-allergen proteins are in the range of 23–98%, with the motif methods giving the best ones at 90–96%. As they are based entirely or partly on local sequence similarity or motifs, these methods are less effective for novel allergen proteins that have no sequence similarity to known proteins. Moreover, recognition of an allergen depends on the properties of the whole protein (Malandain, 2004), which is manifested by the correlation of allergenicity with such properties as

Table 1
Computational methods explored for the prediction of allergen proteins

Publication and year of report	Method	Dataset	Prediction accuracy	
			Allergen	Non-allergen
PMID: 12611632 (Zorz et al., 2002)	kNN	91 Food allergen, 367 non-allergen	81% (Leave-one-out)	98% (Independent test)
PMID: 12709401 (Stadler and Stadler, 2003)	FAO/WHO method	779 Allergen, 2400 shuffled allergens as non-allergens	98%	Precision: 36.6%
	Motif-based	779 Allergen, 2400 shuffled allergens as non-allergens	100%	Precision: 95.5%
PMID: 14739578 (Soeria-Atmadja et al., 2004)	Linear Gaussian	318 Allergen, 1007 non-allergen	75–90%	47–92%
	Quadratic Gaussian	318 Allergen, 1007 non-allergen	74–95%	23–92%
	kNN	318 Allergen, 1007 non-allergen	73–87%	66–92%
	Linear Gaussian	318 Allergen, 1007 non-allergen	74–91%	34–94%
PMID: 15117757 (Li et al., 2004)	Motif	664 Allergens randomly shuffled allergen as non-allergens	70% (Cross-validation)	Precision 90%
PMID: 15373946 (Fiers et al., 2004)	FAO/WHO method	334 Swissprot allergen	42–75%	77–84%
		632 WHO-IUIS allergen	25–75%	81–91%
		730 Swissprot and WHO-IUIS allergen	25–75%	84–92%

Publications are given by authors and Pubmed ID of the respective paper.

ligand-binding, membranes and lipid interaction, protein stability and mobility, glycosylation, structural repeat and aggregation (Breiteneder and Mills, 2005b). Therefore, some clues might be lost in the local sequence patterns analyzed by these methods. It is thus desirable to explore methods that predict allergen proteins from the properties of a whole protein (Malandain, 2004).

A statistical learning method, support vector machines (SVM), has shown promising capability for predicting proteins that belong to specific functional class from the sequence-derived structural and physicochemical properties of a whole protein (Cai et al., 2003; Han et al., 2004b), which has been tested for predicting novel proteins that have no known similarity proteins (Han et al., 2004a, 2005). Proteins in a functional class are generally of diverse sequences but share some structural and physicochemical features to support the common function. In many respects, allergen proteins possess similar characteristics in that they also share some structural and physicochemical features to support allergenicity (Breiteneder and Mills, 2005b). Therefore, SVM is expected to be equally applicable for predicting allergen proteins from sequence-derived structural and physicochemical properties. This work examines the feasibility of developing such a SVM prediction system, which is evaluated by comparison with the results of other computational methods and by a screening test against all proteins in the Swissprot database (Gasteiger et al., 2001).

It is noted that, while allergenicity may be predicted by classifying allergen proteins into groups of common attributes for sequence analysis or statistical learning, such grouping of allergen proteins is different from that of protein functional families. Allergenicity is a quite different type of functional property since it has no function for the producing organism. Many or most described allergen molecules have such a known original function as enzyme, enzyme inhibitor, and transporter. Their allergenicity appears to be no more than a rather unfortunate side-effect for by-standers without any clear disadvantage or advantage for the producing organism. Thus also from an evolutionary perspective there is not much reason to expect that allergens may be readily classified into functional families. While they indeed can be grouped into families based on their structural and biological similarities, this has possibly more to do with shared features of the immune system of allergic individuals.

2. Methods

2.1. Allergen and non-allergen protein datasets

A total of 1005 allergen proteins are collected from Allergome database (<http://www.allergome.org>), which is the most carefully updated allergen database containing entries of usually high reliability scores (Malandain, 2004). Non-allergen proteins are selected by the following procedure: First, all of the protein families in the Pfam database (Bateman et al., 2002) that contain no single known allergen are selected. Secondly, several representative proteins from each of these selected protein families are chosen as non-allergen proteins based on the criteria that their host species are in constant and widest possible contact with humans of all regions and ethnical groups, or the host species

have been extensively studied for allergenicity. This is to ensure that the selected proteins are most unlikely un-identified allergens. The host species considered in this work include human, bovine, chicken, pear, apple, peanut and some others.

These allergen and non-allergen proteins are divided into separate training, testing and independent evaluation sets by the following procedure. First, proteins were clustered into groups based on their distance in the structural and physicochemical feature-space by using the hierarchical clustering method 25. An upper-limit of the largest separation of 20 was used to for each cluster. One representative protein was randomly selected from each group to form a training set that is sufficiently diverse and broadly distributed in the feature space. One or up to 50% of the remaining proteins in each group were randomly selected to form the testing set. The selected proteins from each group were further checked to ensure that they are distinguished from proteins from other groups. The remaining proteins were used as the independent evaluation set, which are also of reasonable level of diversity. Moreover, an analysis of the “similarity” proteins in each cluster shows that the majority of the proteins in a cluster are non-homologous. Thus, the testing and independent evaluation sets are expected to have certain level of usefulness for performing their task of fine-tuning the parameter of a SVM classification system and for evaluating its prediction performance. There are 554, 222, 229 allergens and 7883, 7869, 6717 non-allergens from 7871 Pfam families in these three datasets, respectively.

2.2. Derivation of structural and physicochemical properties from protein sequence

A feature vector can be constructed for representing the structural and physicochemical properties of a protein based on a well-formulated procedure (Cai et al., 2003; Han et al., 2004a, 2005). Given the sequence of a protein, its amino acid composition and the properties of every constituent amino acid are computed and then used to generate this vector. The computed amino acid properties include hydrophobicity, normalized Van der Waals volume, polarity, polarizability, charge, surface tension, secondary structure and solvent accessibility (Cai et al., 2003).

For each of these properties, amino acids are divided into three groups such that those in a particular group are regarded to have approximately the same property. For instance, amino acids can be divided into hydrophobic (CVLIMFW), neutral (GASTPHY), and polar (RKEDQN) groups. Three descriptors, composition (*C*), transition (*T*), and distribution (*D*), are introduced to describe global composition of each of these properties. *C* is the number of amino acids of a particular property (such as hydrophobicity) divided by the total number of amino acids in a protein sequence. *T* characterizes the percent frequency with which amino acids of a particular property is followed by amino acids of a different property. *D* measures the chain length within which the first, 25%, 50%, 75% and 100% of the amino acids of a particular property is located, respectively. Overall, there are 21 elements representing these three descriptors: 3 for *C*, 3 for *T* and 15 for *D*. The feature vector of a protein is constructed by

combining the 21 elements of all of these properties and the 20 elements of amino acid composition in sequential order.

2.3. Support vector machines

The algorithms of SVM and its applications to proteins are extensively described in the literature (Burges, 1998; Cai et al., 2003; Han et al., 2004a). Thus only a brief description is given here. A linear SVM constructs a hyperplane that separates two different classes of feature vectors with a maximum margin. One class represents allergens and the other non-allergens. This hyperplane is constructed by finding a vector $\mathbf{w}$ and a parameter b that minimizes $\|\mathbf{w}\|^2$ which satisfies the following conditions: $\mathbf{w} \cdot \mathbf{x}_i + b \geq +1$, for $y_i = +1$ (positive class) and $\mathbf{w} \cdot \mathbf{x}_i + b \leq -1$, for $y_i = -1$ (negative class). Here $\mathbf{x}_i$ is a feature vector, y_i is the group index, $\mathbf{w}$ is a vector normal to the hyperplane, $|b|/\|\mathbf{w}\|$ is the perpendicular distance from the hyperplane to the origin, and $\|\mathbf{w}\|^2$ is the Euclidean norm of $\mathbf{w}$.

A non-linear SVM projects feature vectors into a high dimensional feature space by using a kernel function such as $K(\mathbf{x}_i, \mathbf{x}_j) = e^{-\|\mathbf{x}_i - \mathbf{x}_j\|^2/2\sigma^2}$. The linear SVM procedure is then applied to the feature vectors in this feature space. After the determination of $\mathbf{w}$ and b , a given vector $\mathbf{x}$ can be classified by using $\text{sign}[(\mathbf{w} \cdot \mathbf{x}) + b]$, a positive or negative value indicates that the vector $\mathbf{x}$ belongs to the positive or negative class, respectively.

The performance of SVM has been measured by the sensitivity $SE = TP/(TP + FN)$, specificity $SP = TN/(TN + FP)$, precision $PR = TP/(TP + FP)$ and overall prediction accuracy $P = (TP + TN)/N$, which correspond to the accuracies for allergens, non-allergens and all proteins, respectively. Here TP, TN, FP, and FN are the number of true positive, true negative, false positive, and false negative, respectively, and $N = TP + FN + TN + FP$ is the total number of proteins. In some cases, SE, SP, PR, and P are insufficient to provide a complete assessment of the performance of a discriminative method MMP (Baldi et al., 2000; Provost et al., 1998). Thus the Matthews correlation coefficient (MCC) was used in this work to evaluate the randomness of the prediction. The MCC has been defined as follows:

$$MCC = \frac{TPTN - FPFN}{\sqrt{(TP + FN)(TP + FP)(TN + FP)(TN + FN)}}$$

where $MCC \in [-1, 1]$, a negative value of MCC indicate the disagreement of the prediction and a positive value of MCC indicate the agreement of the prediction. A zero value means the prediction is completely random. The MCC utilizes all four basic elements of the accuracy and it provides a better summary of the prediction performance than the overall accuracy.

3. Results

Table 2 gives the results of our SVM prediction system. Based on the test of the independent evaluation set, it is found that 93.0% of the 229 allergen proteins and 99.9% of 6717 non-allergen proteins are correctly predicted, the estimated precision, overall accuracy and Matthews correlation coefficient is 100%, 99.9%, and 95.6%, respectively, which are comparable to the best accuracies of other methods (Bochner and Busse, 2004; Li et al., 2004; Stadler and Stadler, 2003). Our SVM prediction system can be further evaluated by a screening test against all of the 168,128 proteins in the Swissprot database (Gasteiger et al., 2001) to determine whether the predicted number of allergens is consistent with the estimates from other studies. It is found that 4943 or 2.9% of these are predicted as allergen proteins, which is consistent with the estimated number of 4093 or 4.0% of the 101,602 Swissprot protein entries from a motif-based method (Stadler and Stadler, 2003) and 4768 or 3.5% of the 135,850 Swissprot protein entries from a combined motif and BLAST similarity search method (Li et al., 2004). These results suggest that a SVM method based on structural and physico-chemical properties of a whole protein is useful for predicting allergen proteins at accuracies comparable to the motif-based approaches.

One particular application of our SVM system is for the prediction of novel allergen proteins that have no similarity proteins. To test this capability, two groups of novel allergen proteins are selected. One contains allergen proteins having no single homologous protein in the Swissprot database (Gasteiger et al., 2001) and the other contains allergen proteins having no other homologous allergen protein based on PSI-BLAST sequence analysis (Altschul et al., 1997). A similarity E -value threshold of 0.1 is used for homolog search to ensure maximum exclusion of allergen proteins that have a homolog. As shown in Tables 3 and 4, a total of 18 and 29 allergen proteins are

Table 2

Performance and statistics of datasets of our SVM prediction of allergen and non-allergen proteins in the training, testing and independent evaluation sets

Dataset	Allergen		Non-allergen		Prediction accuracy				
	TP	FN	TN	FP	SE (%)	SP (%)	PR (%)	P (%)	MCC (%)
Training	554	0	7883	0	100	100	100	100	100
Testing	221	1	7869	0	99.5	100	100	99.9	99.9
Independent evaluation	213	16	6714	3	93.0	99.9	98.6	99.7	95.6

The prediction results are given in TP (true positive), FN (false negative), TN (true negative), FP (false positive), sensitivity $SE = TP/(TP + FN)$ (accuracy for allergens), specificity $SP = TN/(TN + FP)$ (accuracy for non-allergens), precision $PR = TP/(TP + FP)$, overall accuracy $P = (TN + TP)/N$ and Matthews correlation coefficient $MCC = (TPTN - FPFN)/\sqrt{(TP + FN)(TP + FP)(TN + FP)(TN + FN)}$. The number of allergens, non-allergens and the total number of proteins in each dataset is $TP + FN$, $TN + FP$ and $N = TP + FN + TN + FP$, respectively.

Table 3

List of allergen proteins that have no single homologous protein in Swissprot database, and the status of SVM prediction

Allergen protein	Protein name	Swissprot accession number	SVM prediction status
Cop c 1	Cop c1 allergen	Q9Y7G3	+
Cop c 5	RCop c5	Q9UW00	+
RCop c7	RCop c7	Q9UVZ9	–
Hev b 5	Major latex allergen Hev b 5	Q39967	+
Blo t 12	Major allergen Blo t 12 precursor	Q17282	+
Alt a1	Alta 1	Q00021	+
Hev b citrate-binding protein	Citrate binding protein precursor	Q39962	+
Sal k 1	Pollen allergen Sal k 1	P83181	+
Dac g 4	Major pollen allergen Dac g 4	P82946	+
Alt a 1	Major allergen Alt a 1 precursor	P79085	+
Mala s 9	Mala s 9 Allergen	O93973	+
Mala s 8	Mala s 8 Allergen	O93972	–
Mala s 7	Mala s 7 Allergen	O93971	+
Asp f 17	IgE-binding protein	O60025	+
Asp f 4	Allergen Asp f 4	O60024	+
Asp f 7	Allergen Asp f 7	O42799	+
Ole e 6	Pollen allergen Ole e 6	O24172	+
Aed a 3	30 kDa salivary gland allergen Aed a 3 precursor	O01949	+

The symbol + and – indicates the protein is predicted as allergen and non-allergen, respectively.

Table 4

List of allergen proteins that have no other homologous allergen protein, and the status of SVM prediction

Allergen name	Protein name	SwissProt accession number	SVM prediction status
Cop c 1	Cop c1 allergen	Q9Y7G3	+
Mala f 4	Major allergenic protein Mal f4	Q9Y750	+
RCop c5	RCop c5	Q9UW00	+
RCop c7	RCop c7	Q9UVZ9	–
Cand a 3	29 kDa IgE-binding protein	Q6YK78	+
Cul n 1	Ribosomal P0 protein	Q9BMP6	+
Lep d alpha-tubulin	Tubulin alpha chain	Q8WQ47	+
Art v 1 (b)	Major pollen allergen Art v 1 precursor	Q84ZX5	–
Arg r 1	Arg r 1 precursor	Q5GQ85	+
Hev b 5	Major latex allergen Hev b 5	Q39967	+
Hev b citrate-binding protein	Citrate binding protein precursor	Q39962	+
Hom s 3	BCL7B protein	Q13845	–
Alt a1	Alt a1	Q00021	+
Api g 3	Chlorophyll a–b binding protein	P92919	+
Cuc m 3	Pathogenesis-related protein;PR-1	P83834	+
Sal k 1	Pollen allergen Sal k 1	P83181	+
Hor v 21	Gamma-hordein 3	P80198	+
Aed a 1	ATP-diphosphatase	P50635	+
Cuc p ascorbate oxidase	L-ascorbate oxidase	P37064	+
Asp n 25	3-Phytase B precursor	P34754	+
Asp n glucoamylase	Glucoamylase precursor	P04064	+
Bos d 8	Alpha-S2 casein precursor	P02663	+
Alt a 2	Allergen r Alt a 2	O94095	+
Mala s 9	Mal s 9 Allergen	O93973	+
Mala s 7	Mal s 7 Allergen	O93971	+
Asp f 17	IgE-binding protein	O60025	+
Asp f 4	Allergen Asp f 4	O60024	+
Asp f 15	Allergen Asp f 15 precursor	O60022	+
Chi k 1	ADP, ATP carrier protein	P31692	+

found in these two groups, respectively, in which 16 and 26 or 88.9% and 89.7% are correctly predicted by our SVM prediction system, respectively. Therefore, our SVM prediction system appears to show reasonably good capability for predicting novel allergen proteins based on the limited number of these proteins tested.

4. Discussion

Several factors may affect the accuracy of our SVM prediction system for the identification of allergen proteins. One is the diversity of protein samples used for training SVM. It is likely that not all possible types of allergen proteins are ade-

quately represented in our training set. This can be improved with the accumulation of more allergen protein data. Not all proteins of the same function have similar structural and chemical features. There are cases in which different functional groups, un-conserved with respect to position in the primary sequence, mediate the same mechanistic role, due to the flexibility at the active site (Todd et al., 2002). This plasticity is unlikely to be sufficiently described by the structural and physicochemical descriptors currently used in our SVM prediction system, which may affect the prediction of some allergen proteins. Additional structural and physicochemical descriptors may need to be introduced to alleviate this problem.

5. Concluding remark

SVM shows certain level of capability for predicting allergen proteins from the sequence-derived structural and physicochemical properties of a whole protein. It is particularly useful for novel allergen proteins that are non-homologous to any known protein. Further improvements in the training set coverage of allergen protein types, allergen sample collections, and SVM algorithm may enable the development of SVM into a practical tool for facilitating the prediction of allergen proteins directly from protein sequence.

References

- Altschul, S., Madden, T., Schaffer, A., Zhang, J., Zhang, Z., Miller, W., Lipman, D., 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucl. Acids Res.* 25, 3389–3402.
- Baldi, P., Brunak, S., Chauvin, Y., Andersen, C.A., Nielsen, H., 2000. Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics* 16, 412–424.
- Bateman, A., Birney, E., Cerruti, L., Durbin, R., Eddy, S., Griffiths-Jones, S., Howe, K., Marshall, M., Sonnhammer, E., 2002. The Pfam protein families database. *Nucl. Acids Res.* 30, 276–280.
- Bochner, B.S., Busse, W.W., 2004. Advances in mechanisms of allergy. *J. Allergy Clin. Immunol.* 113, 868–875.
- Breddehorst, R., David, K., 2001. What establishes a protein as an allergen? *J. Chromatogr. B, Biomed. Sci. Appl.* 756, 33–40.
- Breiteneder, H., Clare Mills, E.N., 2005. Plant food allergens—structural and functional aspects of allergenicity. *Biotechnol. Adv.* 23, 395–399.
- Breiteneder, H., Ebner, C., 2000. Molecular and biochemical classification of plant-derived food allergens. *J. Allergy Clin. Immunol.* 106, 27–36.
- Breiteneder, H., Ebner, C., 2001. Atopic allergens of plant foods. *Curr. Opin. Allergy Clin. Immunol.* 1, 261–267.
- Breiteneder, H., Mills, C., 2005a. Nonspecific lipid-transfer proteins in plant foods and pollens: an important allergen class. *Curr. Opin. Allergy Clin. Immunol.* 5, 275–279.
- Breiteneder, H., Mills, E.N., 2005b. Molecular properties of food allergens. *J. Allergy Clin. Immunol.* 115, 14–23 (quiz 24).
- Brusic, V., Petrovsky, N., Gendel, S., Millot, M., Gigonzac, O., Stelman, S., 2003. Computational tools for the study of allergens. *Allergy* 58, 1083–1092.
- Burges, C., 1998. A tutorial on support vector machines for pattern recognition. *Data Min. Knowl. Disc.* 2, 121–167.
- Cai, C., Han, L., Ji, Z., Chen, X., Chen, Y., 2003. SVM-Prot: web-based support vector machine software for functional classification of a protein from its primary sequence. *Nucl. Acids Res.* 31, 3692–3697.
- Cebra, J.J., Fuhrman, J.A., Griffin, P., Rose, F.V., Schweitzer, P.A., Zimmerman, D., 1984. Changes in specific B cells and the dissemination of the primed state in vivo following antigenic stimulation by different mucosal routes. *Ann. Allergy* 53, 541–549.
- Cookson, W., 2004. The immunogenetics of asthma and eczema: a new focus on the epithelium. *Nat. Rev. Immunol.* 4, 978–988.
- Davis, P.J., Smales, C.M., James, D.C., 2001. How can thermal processing modify the antigenicity of proteins? *Allergy* 56 (Suppl. 67), 56–60.
- Fiers, M.W., Kleter, G.A., Nijland, H., Peijnenburg, A.A., Nap, J.P., van Ham, R.C., 2004. Allermatch, a webtool for the prediction of potential allergenicity according to current FAO/WHO Codex alimentarius guidelines. *BMC Bioinform.* 5, 133.
- Gasteiger, E., Jung, E., Bairoch, A., 2001. SWISS-PROT: connecting biomolecular knowledge via a protein database. *Curr. Issues Mol. Biol.* 3, 47–55.
- Gould, H., Sutton, B., Beavil, A., Beavil, R., McCloskey, N., Coker, H., Fear, D., Smurthwaite, L., 2003. The biology of IGE and the basis of allergic disease. *Ann. Rev. Immunol.* 21, 579–628.
- Han, L., Cai, C., Ji, Z., Cao, Z., Cui, J., Chen, Y., 2004a. Predicting functional family of novel enzymes irrespective of sequence similarity: a statistical learning approach. *Nucl. Acids Res.* 32, 6437–6444.
- Han, L., Cai, C., Ji, Z., Chen, Y., 2005. Prediction of functional class of novel viral proteins by a statistical learning method irrespective of sequence similarity. *Virology* 331, 136–143.
- Han, L., Cai, C., Lo, S., Chung, M., Chen, Y., 2004b. Prediction of RNA-binding proteins from primary sequence by a support vector machine approach. *RNA* 10, 355–368.
- Hileman, R., Silvanovich, A., Goodman, R., Rice, E., Holleschak, G., Astwood, J., Hefle, S., 2002. Bioinformatic methods for allergenicity assessment using a comprehensive allergen database. *Int. Arch. Allergy Immunol.* 128, 280–291.
- Huby, R.D., Dearman, R.J., Kimber, I., 2000. Why are some proteins allergens? *Toxicol. Sci.* 55, 235–246.
- Ivanciuc, O., Mathura, V., Midoro-Horiuti, T., Braun, W., Goldblum, R., Schein, C., 2003. Detecting potential IgE-reactive sites on food proteins using a sequence and structure database, SDAP-food. *J. Agric. Food Chem.* 51, 4830–4837.
- Kleter, G., Peijnenburg, A., 2002. Screening of transgenic proteins expressed in transgenic food crops for the presence of short amino acid sequences identical to potential, IgE-binding linear epitopes of allergens. *BMC Struct. Biol.* 2, 8.
- Lack, G., 2004. New developments in food allergy: old questions remain. *J. Allergy Clin. Immunol.* 114, 127–130.
- Li, K.B., Issac, P., Krishnan, A., 2004. Predicting allergenic proteins using wavelet transform. *Bioinformatics* 20, 2572–2578.
- Malandain, H., 2004. Basic immunology, allergen prediction, and bioinformatics. *Allergy* 59, 1011–1012.
- Mills, E.N., Jenkins, J.A., Alcocer, M.J., Shewry, P.R., 2004. Structural, biological, and evolutionary relationships of plant food allergens sensitizing via the gastrointestinal tract. *Crit. Rev. Food Sci. Nutr.* 44, 379–407.
- Ono, S., 2000. Molecular genetics of allergic diseases. *Ann. Rev. Immunol.* 18, 347–366.
- Provost, F., Fawcett, T., Kohavi, R., 1998. The case against accuracy estimation for comparing induction algorithms. In: *Proceedings of the 15th International Conference on Machine Learning*, Morgan Kaufmann, San Francisco, CA, pp. 445–453.
- Sampson, H., 2004. Update on food allergy. *J. Allergy Clin. Immunol.* 113, 805–819.
- Sicherer, S., Munoz-Furlong, A., Sampson, H., 2004. Prevalence of seafood allergy in the United States determined by a random telephone survey. *J. Allergy Clin. Immunol.* 114, 159–165.
- Soeria-Atmadja, D., Zorzet, A., Gustafsson, M., Hammerling, U., 2004. Statistical evaluation of local alignment features predicting allergenicity using supervised classification algorithms. *Int. Arch. Allergy Immunol.* 133, 101–112.
- Stadler, M.B., Stadler, B.M., 2003. Allergenicity prediction by protein sequence. *FASEB J.* 17, 1141–1143.
- Stewart, G.T., 1973. Allergy to penicillin and related antibiotics: antigenic and immunochemical mechanism. *Annu. Rev. Pharmacol.* 13, 309–324.
- Sutton, B., Gould, H., 1993. The human IgE network. *Nature* 366, 421–428.

- Todd, A., Orengo, C., Thornton, J., 2002. Plasticity of enzyme active sites. *Trends Biochem. Sci.* 27, 419–426.
- van Wijk, F., Nierkens, S., Hassing, I., Feijen, M., Koppelman, S.J., de Jong, G.A., Pieters, R., Knippels, L.M., 2005. The effect of the food matrix on in vivo immune responses to purified peanut allergens. *Toxicol. Sci.* 86, 333–341.
- Venge, J., Lampinen, M., Hakansson, L., Rak, S., Venge, P., 1996. Identification of IL-5 and RANTES as the major eosinophil chemoattractants in the asthmatic lung. *J. Allergy Clin. Immunol.* 97, 1110–1115.
- Wal, J.M., 2001. Structure and function of milk allergens. *Allergy* 56 (Suppl. 67), 35–38.
- Zorzet, A., Gustafsson, M., Hammerling, U., 2002. Prediction of food protein allergenicity: a bioinformatic learning systems approach. *In Silico Biol.* 2, 525–534.