Published online August 23, 2006

Nucleic Acids Research, 2006, Vol. 34, No. 13 ***3779–3793***
doi:10.1093/nar/gkl467

Computational detection of allergenic proteins attains a new level of accuracy with *in silico* variable-length peptide extraction and machine learning

D. Soeria-Atmadja, T. Lundell, M. G. Gustafsson[1,2,]* and U. Hammerling*

Division of Toxicology, National Food Administration, PO Box 622, SE-751 26 Uppsala, Sweden,
[1]Department of Engineering Sciences, Uppsala University, PO Box 534, SE-751 21 Uppsala, Sweden and
[2]Department of Genetics and Pathology, Uppsala University, Rudbeck Laboratory, SE-751 85 Uppsala, Sweden

Received April 4, 2006; Revised and Accepted June 20, 2006

ABSTRACT

The placing of novel or new-in-the-context proteins on the market, appearing in genetically modified foods, certain bio-pharmaceuticals and some household products leads to human exposure to proteins that may elicit allergic responses. Accurate methods to detect allergens are therefore necessary to ensure consumer/patient safety. We demonstrate that it is possible to reach a new level of accuracy in computational detection of allergenic proteins by presenting a novel detector, Detection based on Filtered Length-adjusted Allergen Peptides (DFLAP). The DFLAP algorithm extracts variable length allergen sequence fragments and employs modern machine learning techniques in the form of a support vector machine. In particular, this new detector shows hitherto unmatched specificity when challenged to the Swiss-Prot repository without appreciable loss of sensitivity. DFLAP is also the first reported detector that successfully discriminates between allergens and non-allergens occurring in protein families known to hold both categories. Allergenicity assessment for specific protein sequences of interest using DFLAP is possible via ulfh@slv.se.

INTRODUCTION

Allergic diseases are characterized by immunologic responses against otherwise innocuous substances, typically proteins (1–4). Allergy grows steadily and may now affect >20% of the population in the Western hemisphere. The clinical manifestations can involve any one of various different symptoms such as asthma, rhinitis, rhinoconjunctivitis, eczema, contact dermatitis, angioedema, abdominal pain and anaphylaxis, the latter being a potentially life-threatening condition (5,6). An immediate allergic response (type I hypersensitivity, as opposed to delayed-type, cell-mediated allergic reactions) includes the preferential synthesis and secretion of immunoglobulin E (IgE) molecules, which subsequently anchor efficiently to high-affinity receptors on tissue mast cells and basophilic granulocytes. Contact with proteins that can simultaneously bind to at least two cell surface-attached IgE-molecules triggers the secretion of inflammatory mediators and cytokines (3,4). It is essential to distinguish between complete and incomplete protein allergens, i.e. those which can educate the immune system (sensitization) to a full response and those which only have the ability to trigger release of mediators through cross-reactive IgE binding, respectively (7–10). Many allergens appear to cluster into relatively few protein families (7,11–13). Nonetheless, most members of such protein families seem to be devoid of allergenic properties (14).

A number of immunochemical, biochemical and immunological methods for the identification proteins with a potential to cause of type-I hypersensitivity reactions have emerged and evolved over time, notably IgE immunosorbent assays using patient sera, human skin prick tests and basophil histamine release as well as various animal models, the first two assay methods also being commonly used clinical diagnostic modalities (15–17). In recent years, however, bioinformatics tests for allergenicity have become increasingly visible in the literature. Largely, this direction has emerged in response to a need for a relatively expedient method to screen for potential protein allergens owing to concern over the possibility of unpremeditated introduction of allergen-encoding transgenes into genetically modified (GM) food crops (18,19). In 1996, the ILSI/IFBC presented a decision-tree for safety assessment of GM foods, which encompasses several principally dissimilar testing methods including an introductory amino acid sequence comparison for xenoproteins, obtained from sources with known allergenic potential,

*To whom correspondence should be addressed. Email: ulfh@slv.se
*Correspondence may also be addressed to M. G. Gustafsson. Tel: +46 18 4713229; Fax: +46 18 555096; Email: mg@angstrom.uu.se
Present address:
M. G. Gustafsson, Department of Medical Sciences, Uppsala University, Uppsala University Hospital, SE-751 85 Uppsala, Sweden

to allergen sequences (20). Several years later, an FAO/WHO consultation on allergenic foods presented a revised scheme, in which partly similar bioinformatics testing is a mandatory introductive step irrespective of transgene origin. This protocol prescribes a two-part procedure wherein a protein is assigned as an allergen by either a match of six consecutive amino acids or an identity of >35%, across an 80 amino acid window, in both cases to any documented protein allergen (21). Subsequently, the *Codex Alimentarius* summoned a panel of regulatory experts to review the FAO/WHO recommendations. The final guidance prescribes that the peptide size in contiguous amino acid searches be based on a scientifically justified rationale (22). There are several websites that offer allergenicity testing of query amino acid sequences according to the FAO/WHO *in silico* protocol alone or as part of several interrogation formats: AllerPredict, (http://research.i2r.a-star.edu.sg/Templar/DB/Allergen/Predict/Predict.html), Structural Database of Allergenic Proteins (SDAP) (23), (http://fermi.utmb.edu/SDAP/sdap_who.html), AllerMatch (24), (http://www.allermatch.org/) and Allergen Database for Food Safety (ADFS) (25), (http://allergen.nihs.go.jp/ADFS/).

Various additional bioinformatics testing procedures founded on inter-allergen similarity have been published, such as those relying on the FASTA algorithm and various stretches of identical amino acid matches (26–31) or on automated motif discovery, either as standalone methods or in combination with conventional sequence similarity search (32,33). A different course, which takes advantage of available data on IgE epitopes and amino acid descriptors, is also described (23,34). Recently, we reported an *in silico* detector of potential protein allergens based on a novel principle wherein peptides enriched for allergenic features are obtained through a special selection procedure, involving sequence comparison between peptides of allergens and non-allergens (35). The detection system is named Detection based on Automated Selection of Allergen-Representative Peptides (DASARP).

Although many promising computational methods for detection of protein allergens have been reported, none of them allows successful discrimination between allergens and non-allergens within particular protein families such as the tropomyosin family. Moreover, false alarm rates of methods described, when tested on non-allergens, are much higher than desired and expected. In this article we demonstrate for the first time that new computational approaches and databases enable a much higher level of performance. In particular, we show that a computational course allowing extraction and employment of variable length peptides, in conjunction with modern machine learning techniques, offers a new level of accuracy compared with earlier attempts. The resulting detector, denoted Detection based on Filtered Length-adjusted Allergen Peptides (DFLAP), derives a flexible number of peptides of variable lengths per allergen as controlled by an adjustable threshold parameter. The obtained fragments are designated Filtered Length-adjusted Allergen Peptides (FLAPs). In analogy with other allergenicity prediction methods based on sequence similarity to known allergens, DFLAP is most suitable for detection of cross-reactive allergens, although otherwise allergenic proteins may also be correctly detected.

DFLAP operates expediently enabling evaluation of the method with the entire Swiss-Prot database (36). The detection performance results are shown to be roughly equal to that of ILSI/IFBC and DASARP mentioned above. However, DFLAP assigns a much lower fraction of the Swiss-Prot database as being allergens than any hitherto published report of comparable detection rate, thus demonstrating much higher specificity. DFLAP is also the first *in silico* detector reported capable of distinguishing allergens from non-allergens within protein families, which is most conspicuous in the case of tropomyosins.

MATERIALS AND METHODS

Outline of DFLAP method

FLAPs were extracted from an allergen database through a specially designed comparison with a non-allergen dataset and, when applicable, a subsequent concatenation procedure (see below for details on data repositories). This procedure, described in detail below, is designated Computerized Peptide Filtration and Aggregation (CPFA). Downstream of the CPFA step, the DFLAP algorithm may be summarized as follows: Amino acid sequences of query proteins are compared with the extracted FLAPs. Based on the resulting list of similarity score values for each protein, a support vector machine (SVM) is trained to decide whether there is sufficient similarity to assign the query protein as an allergen. A point-wise summary of the method is shown in Figure 1 and a detailed description of datasets and the different sub-procedures of DFLAP can be found below.

Computerized peptide filtration and aggregation

First, each allergen was segmented into amino acids of length $l_{\min}$ through a sliding window procedure. Each derived unit was then aligned against the non-allergen dataset and the top alignment score for each peptide was stored. Thus, for each allergen a filter score vector V_{FS} of length $L_i - l_{\min} + 1$ was stored, where L_i is the length of allergen i. Each peptide score in V_{FS} was thereafter compared with a FLAP threshold in order to transform V_{FS} into a discrete binary-valued vector. Whenever a peptide's corresponding value in V_{FS} was below the FLAP threshold, the corresponding value in the binary vector was set to one, i.e. the peptide was deemed dissimilar enough to any known non-allergen. Two or more overlapping peptides, defined as consecutive units in the binary V_{FS}, were concatenated into a single longer fragment. Since the extracted peptides have variable length, depending on whether they are non-modified or concatenated, they were collectively referred to as FLAPs or simply the FLAP set.

Alignment of allergen peptides to the non-allergen database was conducted by virtue of our own tailor-made MATLAB toolbox that allows fast alignment between a complete amino acid sequence and multiple shorter peptides. The following parameter settings were applied: BLOSUM62 substitution matrix and $-11/-1$ gap penalties.

Generation and extraction of features

Amino acid sequences, either applied to design a classifier by means of supervised learning or used as input queries, were aligned against all peptides in the FLAP set. Each sequence generated a characteristic assembly of alignment score

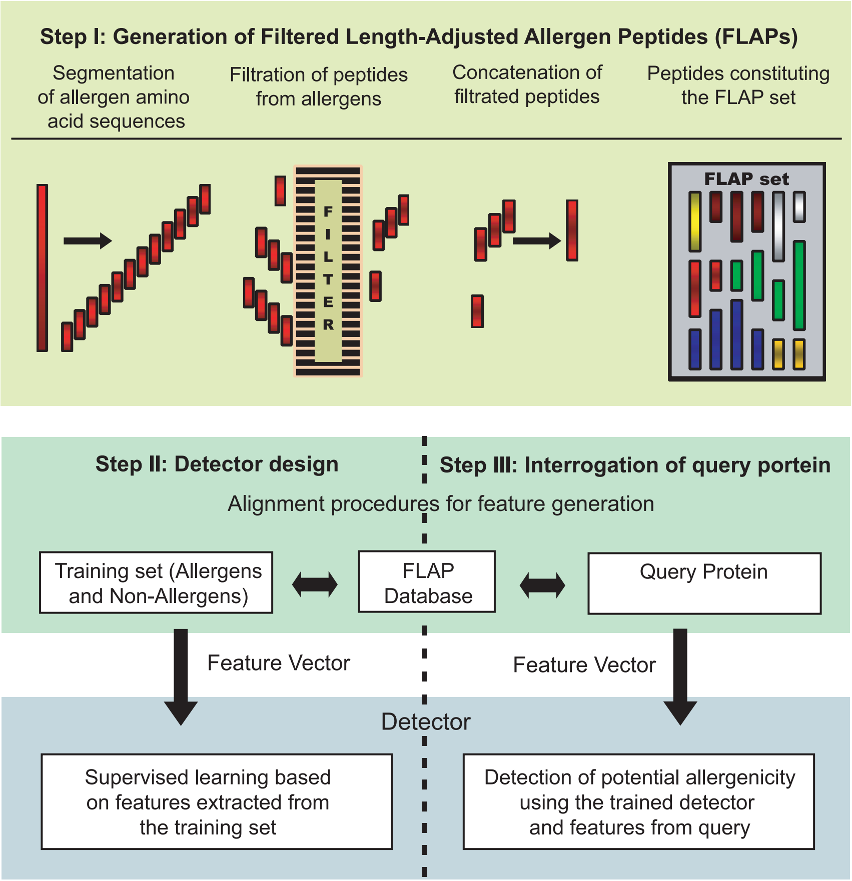
Step I: Generation of Filtered Length-Adjusted Allergen Peptides (FLAPs)
Segmentation of allergen amino acid sequences
Filtration of peptides from allergens
Concatenation of filtrated peptides
Peptides constituting the FLAP set
FILTER
FLAP set
Step II: Detector design
Step III: Interrogation of query portein
Alignment procedures for feature generation
Training set (Allergens and Non-Allergens)
FLAP Database
Query Protein
Feature Vector
Feature Vector
Detector
Supervised learning based on features extracted from the training set
Detection of potential allergenicity using the trained detector and features from query

Figure 1. Outline of design and function of the DFLAP algorithm. (I) The allergen amino acid sequences are segmented into overlapping peptides and are subsequently compared with all sequences in the non-allergen set. The rational is that peptides, with high similarity to any non-allergen sequence, are likely to be structurally/functionally unrelated to allergenicity. Conversely, peptides lacking appreciable similarity with non-allergen sequences are potentially important to allergic reactions, broadly defined. These peptides (after concatenation of directly overlapping peptides) are stored in a special catalogue designated Filtered Length-adjusted Allergen Peptide (FLAP) set. Thus, the non-allergen amino acid sequence set can be regarded as a filter wherein only peptides dissimilar to non-allergens are allowed to pass. (II) Feature vectors, based on the alignment scores between training amino acid sequences (both allergen and non-allergen) against the FLAP set, are thereafter allowed to educate a supervised learning algorithm. This process trains the algorithm to determine, in a quantitative manner, the level of similarity to the FLAP set, which is required for a protein to be assigned as an allergen. (III) The educated system allows for interrogation by any query amino acid sequence with respect to allergen potential, essentially as described in step II. If sufficient similarity to the FLAPs is found, the query sequence is assigned as an allergen. In this step, the trained detector quantifies what 'sufficient similarity' means.

values, one for each FLAP. Such sets of score values were each refined and reduced, initially by sorting the score values in descending order and subsequently by retaining only the n largest values. Consequently, the final feature vector for a given amino acid sequence consisted of its n top similarity scores against the FLAP set, assorted in descending order. The alignment of amino acid sequences to the FLAP set was conducted by means of the aforementioned in-house alignment algorithm.

Designing a classification algorithm using supervised learning

A linear kernel SVM (37,38) classification algorithm was trained using positive (allergen) and negative (non-allergen) training samples, represented by their respective feature vectors. Allergens were drawn from the same set as those used to generate the FLAPs; non-allergen training samples were obtained by randomly sampling proteins from Swiss-Prot (followed by removal of possible allergens). To provide for a reasonable coverage of protein diversity, the non-allergen training samples were twice as many as the allergens.

Detection of potential allergenicity using the DFLAP system

For any given query sequence, a feature vector was generated using the already described alignment against the FLAP set followed by sorting and selection of the top n features in descending order. Each feature vector was then presented to the trained SVM for an immediate decision.

Performance evaluation and comparison to other computational methods

Three separate tests for evaluation of DFLAP performance were conducted, as presented below and as outlined in Table 1 (DFLAP detectors were built on the most promising

Table 1. Outline of tests conducted

Test type	Detection system	Sequences used for generation of FLAPs (Allergens/non-allergens)*	Training sequences (Allergens/non-allergens)*	Test sequences (Allergens/non-allergens)*
Parameter selection (3-fold CV)	DFLAP	333[a]/52081[b]	333[a]/666[c]	167[d]/339 (334[c]+5[e])
Assessment of sensitivity (holdout)	FAO/WHO	—	500[a]	262[d] (168, 141, 116, 99)[f]
	ILSI/IFBC	—	500[a]	262[d] (168, 141, 116, 99)[f]
	DASARP	—	500[a]	262[d] (168, 141, 116, 99)[f]
	DFLAP**	500[a]/52081[b]	500[a]/1000[c]	262[d] (168, 141, 116, 99)[f]
Assessment of intra-family discrimination (holdout)	FAO/WHO	—	697[a]	65[d]/193[g]
	ILSI/IFBC	—	697[a]	65[d]/193[g]
	DASARP	—	697[a]	65[d]/193[g]
	DFLAP**	697[a]/52081[b]	697[a]/1394[c]	65[d]/193[g]
Assessment of specificity (holdout)	FAO/WHO	—	762	164970[h]
	ILSI/IFBC	—	762	164970[h]
	DASARP	—	762	164970[h]
	DFLAP**	762/52081[b]	762/1524	164970[h]

*All datasets are publicly available on http://www.slv.se/templates/SLV_Page.aspx?id=14772.

**The parameter setting was $l_{\min} = 22$, FLAP threshold = 48, $n = 4$ and $C = 100$.

[a] Subsets of the total amount (762) of allergens used to train each test method in the different evaluation procedures. In the case of DFLAP these subsets were initially also used to generate of FLAPs.

[b] Non-allergen filter set used in the Computerized Peptide Filtration and Aggregation (CPFA).

[c] Subsets of the total amount (1524) of the sequences referred to as Swiss-Prot non-allergens in Materials and Methods, used for SVM training (and testing in the parameter selection procedure) in the DFLAP method.

[d] Subsets of the total amount (762) of allergens used to test each test method in the different evaluation procedures.

[e] Five tropomyosins used to measure specificity in the evaluation of DFLAP parameters.

[f] The four numbers corresponds to different levels of maximal sequence identity between training and test set (95, 90, 85 and 80%), respectively.

[g] Presumed non-allergens from tropomyosins, profilins and parvalbumins.

[h] Swiss-Prot, release 45.3.

parameters, as identified in the 3-fold CV procedure for parameter selection described further below):

- *Assessment of sensitivity using external holdout test sets with different degrees of homology*: First, a holdout dataset of 262 allergens was randomly selected from the total allergen dataset (762 amino acid sequences) to enable unbiased performance evaluation of the designed SVM classifier. Thereafter, a DFLAP detector was built using a training set of the remaining 500 allergens together with 1000 non-allergens. Finally, a performance estimate of the designed DFLAP was obtained by presenting it to samples in the external holdout test set. In order to study the effect of a gradually decreasing sequence homology, the test set was reduced step-wise so that none of the remaining sequences shared more than a predefined degree of sequence similarity, to each other as well as to the training set. Sequence similarity levels of 95, 90, 85 and 80%, according to FASTA3 (BLOSUM50, gap penalties −12/−2), resulted in test datasets containing 168, 141, 116 and 99 allergens, respectively. The full-sized (262 entities) set and the four reduced test sets were each evaluated separately.
- *Evaluation of DFLAP specificity using external test examples from three protein families:* A DFLAP designed with all allergens available except for 70% of the sequences from three protein families—parvalbumins, profilins and tropomyosins—was built. Subsequently, its specificity was estimated by means of an exclusive test based on only the remaining allergens and plausible non-allergens of these three protein families. The test datasets contained 13 allergen and 121 vertebrate tropomyosins, 43 allergen and 39 animal profilins, as well as 9 allergen and 33 mammalian parvalbumins, respectively, all being presumptive non-allergens, whereas the training set contained 762 − 13 − 43 − 9 = 697 allergens and 2 × 697 = 1394 non-allergens.
- *Estimation of the DFLAP specificity using the entire Swiss-Prot database*. First, a DFLAP design founded on all 762 allergens available and 762 × 2 =1534 non-allergens was performed. All these allergens served to construct a Filtered Allergen Peptide set that subsequently was employed to extract features vectors for all the 762 allergens and 1534 non-allergens. Finally a SVM classifier was designed using the feature vectors generated and then applied to detect all allergens in the entire Swiss-Prot database.

For comparison the bioinformatics schemes ILSI/IFBC (20), FAO/WHO (21) and DASARP (35) were also evaluated in the three above listed experiments. As already mentioned, the FAO/WHO procedure assigns an amino acid sequence as potentially allergenic either if a local sequence alignment returns 35% similarity over any segment of 80 residues to a known allergen or if there is an exact match to a peptide of length six in an allergen. In benchmarks conducted here the FAO/WHO-prescribed alignment procedure was applied to both test schemes. The ILSI/IFBC scheme involves a more vaguely defined alignment-type comparison as well as a search for an identical peptide match, but of length eight residues replacing six. Notably, the selection of a criterion based on eight residues partly agrees with the aforementioned suggestion by *Codex Alimentarius* to consider matching lengths between 6 and 8 residues.

The tested DASARP system was created using the same peptide length (24), number of peptides per allergen (5) and scoring scheme (sum of the two highest scores using an allergenicity threshold at 5.51) as reported earlier (35). For computational reasons the generation of Allergen-Representative Peptides (ARPs) was, however, slightly modified. As an alternative to comparing each allergen peptide to each sliding-window peptide of the non-allergens used for filtration, they were aligned to the *entire* amino acid sequences in the non-allergen dataset. Apart from being computationally much faster relative to individual peptide–peptide comparisons, the use of alignment permits inclusion of gaps as opposed to the comparison algorithm proposed earlier.

Parameter selection using a 3-fold CV loop

An outline of parameter selection is presented in Table 1. The 500 allergens and 1000 presumed non-allergens used to train the DFLAP in the sensitivity evaluation procedure were first allocated together with five vertebrate (and thereby presumably non-allergenic) tropomyosins for parameter selection. The former two separate sets of amino acid sequences, respectively, were randomly split into three subsets (of roughly equal size) used in a 3-fold cross validation (CV). In each CV iteration, 166/167 allergens and 333/334 non-allergens were set aside for test together with five vertebrate tropomyosins (regarded as non-allergens). The remaining allergens were first used to create FLAPs based on the algorithm for computational filtration. Subsequently, together with the remaining non-allergens, the created FLAPs were employed in the generation of features for the training procedure of the SVM algorithm (see above). Each parameter setting resulted in a unique classifier evaluated with the remaining 167 allergen and 334 non-allergen test sequences. The fraction of correctly detected allergens and the fraction of erroneously detected non-allergens in general (false alarms) and non-allergen tropomyosins (tropomyosin false alarms) were recorded for each classifier in each CV iteration. The estimated detection rate and false alarm rates were, for each parameter setting, finally obtained as averages over the three CV iterations. Datasets used in each of the CV iterations are publicly available on http://www.slv.se/templates/SLV_Page.aspx?id=14772.

The four different parameters in the parameter selection procedure can be divided into two sub-categories, as outlined below: three of them specify generation and selection of feature parameters, whereas the fourth specifies the SVM algorithm.

Feature parameters.

- *Minimal peptide length ($l_{\min}$):* The segment (peptide) length of each allergen prior to filtration. Lengths of 10–24, with incremental steps of two amino acid residues, were tested.
- *Retention level (degree of filtration):* The fraction of overlapping sliding-window peptides discarded in the filtration process. The FLAP threshold, to which an allergen peptide's maximum alignment score in the non-allergen sequence database is compared, dictates the retention level. A maximum score below this threshold means that the peptide is sufficiently dissimilar to the non-allergens to qualify as a FLAP. FLAP thresholds were varied step-wise to accomplish average retention levels for the allergens of roughly 45, 55, 65 and 75% (four thresholds for each $l_{\min}$).
- *Number of alignment scores (n):* Query amino acid sequences were aligned against all peptides in the FLAP set, thus generating a fingerprint of alignment scores for each sequence; one alignment value for each FLAP. Such data were sorted to obtain gradually declining scores. A simple decision was then performed in which the top n alignment scores (best matches) were extracted as features. Thus, a feature vector for a given amino acid sequence consists of its top similarity scores against the FLAP set. Tested values of n were 1, 2, 3, 4 and 5.

SVM algorithm parameter.

- *Value of cost parameter (C):* In conformity with other commonly used supervised classifiers (k-nearest neighbour, multilayer perceptron artificial neural networks, decision trees etc), the SVM classifier employed here consists of two integrated parts.

The first part comprises an algorithm (computer program) that can accommodate sets of numerical values as an input and, in response to this feeding, provide outputs in the form of discrete decisions, which assigns each such input as belonging to one among several classes. This process always relies on several adjustable parameters that influence the output decision. To simplify illustration of the discriminating part of an SVM, let us consider the special case where each input list (x_1, x_2) consists of only two score values x_1 and x_2. In this case, an observation (x_1, x_2) may be interpreted as a point in a two-dimensional (2D) representation. The decision made by the SVM simply involves the determination of which side of a pre-defined discrimination line the observation point (x_1, x_2) is situated and a phrasing of this judgment as a succinct output.

The second part of a supervised classifier encompasses a tailor-made algorithm that tunes the adjustable parameters of the discriminating part in order to achieve good performance. A standard objective is to perform well on a set of pre-labelled training examples available from a human supervisor. In the special 2D case considered here for illustrative purposes, this corresponds to an adjustment of the slope and position of the pre-defined discrimination line. The characteristic feature of SVM tuning, relative to other linear classifiers, is the explicit aim to obtain a large 'margin' (geometric distance) between the decision line and the most difficult training examples. In practice, SVM aims at placing the discriminating line in the centre of a training dataset, thereby separating all the examples from the two classes. The basic rationale for this choice is to obtain a robust tuning procedure, involving little influence of noise and errors in the training examples on the final discriminating line.

The SVM tuning procedure discussed above, aiming at a large 'margin', is intuitive but unfortunately not applicable to the common situation where no discriminating line that perfectly separates the training examples exists. In real applications, this hurdle is confronted by means of an additional tuning parameter C that allows adjustment of the discrimination line with the explicit aim to obtain a compromise with relatively few errors among the training examples, while still retaining large margins to those classified correctly. Formally, this parameter specifies the penalty assigned to misclassified training examples. By selection from a range of different values of C it is possible to find a discrimination line that has large 'margins' to the most difficult but correctly classified examples while, at the same time, makes few misclassification errors. More formally, a larger value of C increases the penalty for misclassification errors. A suitable value of C can be found from CV or from tests using external test examples.

In summary, a key issue in SVM tuning is to find a balance between good performance on the particular dataset available and the risk of over fit to misleading chance correlations in that particular dataset. An over fit usually results in an over-optimistic performance on the training examples and, consequently, relatively poor performance on completely new examples. Notably, owing to the particular penalty function used in SVM tuning, a successfully tuned SVM classifier may range from few but large training/test errors to many but relatively small training/test errors. Owing to computational limitations, in our work the following five different values of C were evaluated: 0.1, 1, 10, 100 and 1000.

Implementation

All computations were performed in the MATLAB™ programming environment (The MathWorks Inc., Natick, MA), except for those procedures involving alignment between complete amino acid sequences, which were performed using the FASTA3 program (39). Apart from the core program, the MathWorks Statistics toolbox was used, as well as the OSU SVM Classifier Matlab Toolbox by J. Ma, Y. Zhao and S. Anhalt, which is based on LIBSVM version 2.33 [C.-C. Changand and C.-J. Lin (2001) LIBSVM: a library for SVMs]. Moreover, a tailor made toolbox was created that allows a BLAST-like alignment algorithm for fast alignment between a complete amino acid sequence and multiple shorter peptides (e.g. FLAPs).

Bayesian confidence intervals

A 95% Bayesian confidence (BC) or credibility interval $[a,b]$ for an unknown misclassification rate q is any interval for the value of q which contains 95% of the probability mass of the conditional posterior probability density distribution $p(q\,|\,k_t,N_t)$ where k_t is the number of errors made using N_t test examples. The posterior $p(q\,|\,k_t,N_t)$ is calculated using Bayes theorem as $p(q\,|\,k_t,N_t) = P(k_t\,|\,q,N_t)p(q)/P(k_t\,|\,N_t)$ using a uniform prior $p(q) = 1$ on the interval [0,1] (37). Usually the highest posterior density interval is employed which is the shortest among all the possible BC intervals. A BC interval relies on Bayesian inference viewed as extended logic and is therefore not identical to a classical confidence interval that relies on a frequentistic definition of probabilities (40). A BC interval reflects our current knowledge about the unknown performance q and equals a classical confidence interval in cases where the distribution $p(q)$ of true performances is uniformly distributed on the unit interval [0,1]. As a consequence, a Bayesian 95% confidence interval may be interpreted as a conservative form of a 95% classical confidence interval that covers the true value with a probability >95%.

Datasets

Allergen database. The in-house allergen database contained 762 amino acid sequences. The sequences were mined from the following six publicly available databases: Allergen list maintained by IUIS Allergen Nomenclature Sub-Committee (41) (http://www.allergen.org/List.htm), Farrp (29) (http://www.allergenonline.com), The Allergen Database (http://allergen.csl.gov.uk), The Allergen Sequence Database (42) (http://www.iit.edu/~sgendel/fa.htm), The Protall Database (http://www.ifr.bbsrc.ac.uk/protall) and Allergome (http://www.allergome.org). Prior to deposit into our in-house repository the records were manually inspected for documentation on allergenicity (preferably published reports). Records without or with poor such documentation were dismissed. In addition, sequences occurring as fragments shorter than 100 amino acids were discarded to reduce the risk of incorporating truncated protein allergens without sensitizing or cross-reactive regions. The total dataset and the different subsets hereof (used for SVM training/testing in procedures for both parameter selection and performance evaluation) are publicly available on http://www.slv.se/templates/SLV_Page.aspx?id=14772.

Non-allergen dataset for FLAP extraction. Corner stones of DFLAP are extraction and application of allergen representative FLAPs, i.e. peptides of different lengths that rarely occur in non-allergens. The computerized filtration and aggregation algorithm designed for this purpose employs two databases, one holding allergen sequences only (as described above) and another constructed to exclusively contain non-allergens. Clearly, presence of allergens in the latter set may cause elimination of many important allergen representative peptides, thereby reducing detection performance. Accordingly, dedicated data resources were targeted to ensure a minimum of contamination in this non-allergen repository. Two main sources were used:

- *The human proteome*. This proteome is large and should, for obvious reasons, contain a minimum of allergens.
- *Skin prick preparations*. These formulations are used in clinical settings for the specific diagnosis allergic responses in atopic patients. Their established usage in skin prick tests as whole extracts provides the rational for their inclusion in the non-allergen database after removal of reported allergens.

In those cases where a non-fractioned offending organism is included in the skin prick solution (as opposed to just parts such as fur), it follows that the entire organism's proteome has been subjected to the patient. The entire part of the organism's proteome, available as characterized amino acid sequences, was downloaded. Subsequent to depleting them of all known allergens and their respective isoallergens the remaining respective fractions, with a history of exposure to many atopics, were assumed pure of contaminants and thereby judged suitable for inclusion in the non-allergen database. For these sources, the following criteria were applied to all sequences entering the database:

- The textual description of the protein must not contain the word 'allergen'. For the human proteome, an annotation as 'antigen' was neither allowed to pass.

- The protein must have no higher sequence identity than 67% with any of the allergens (the definition of isoallergenicity). Because we assume the risk for contamination of human allergens be appreciably lower than that of other proteomes, the requirement for disparity was relaxed to 80% sequence identity.
- The length of the protein must be longer than or equal to 50. This filters out a large number of redundant protein fragments that are already present in the database in the form of complete proteins.

The final dataset of totally 52 081 sequences, which was used to generate FLAPs (Figure 1) in the evaluation procedures outlined in Table 1, consisted of sequences from the following organisms and is available upon request:

- *Homo sapiens* (human), 50 957 sequences
- *Aspergillus fumigatus* (fungus from mold), 503 sequences
- *Candida albicans* (fungus), 619 sequences
- *Dermatophagoides pteronyssinus* (house dust mite), 2 sequences

Swiss-Prot database. The Swiss-Prot database served as a resource for validation of specificity. For this project the UniProt FASTA release 3.3, which corresponds to Swiss-Prot release 45.3, was used. This database, containing 164 970 sequences, was obtained at http://www.ebi.ac.uk/FTP/.

Swiss-Prot non-allergens for DFLAP training. Apart from the allergens, a 'non-allergen' dataset was created by random sampling from Swiss-Prot to be used for training of the SVM algorithm in DFLAP. In order to minimize contamination, the amino acid sequences had to be at least 50 amino acid residues long and were not allowed to be identical or share high sequence similarity to any of the allergens. The last criterion was controlled by performing simple alignments between the allergens and the sampled excerpt, using FASTA3 (BLOSUM50 and −12/−2 gap penalties). If the Smith–Waterman score used surpassed the limit of 100 the sequence was assigned as too uncertain to be kept as a non-allergen. The resulting dataset, was twice the size of the total number of allergens, i.e. $762 \times 2 = 1524$, and different subsets of this were used for SVM training in the parameter selection procedure as well as in the three different performance evaluation tests (Table 1). The total set of 1524 sequences as well as the different subsets are publicly available on http://www.slv.se/templates/SLV_Page.aspx?id=14772.

Allergen and presumably non-allergen members in different protein families. In order to further evaluate specificity of the algorithm, a DFLAP was built using all allergens except for 70% of those that belongs to the following protein families: parvalbumins, profilins and tropomyosins. Moreover, probable non-allergens from these protein families were also collected and tested for potential allergenicity and consisted of the following datasets: 121 vertebrate tropomyosins, 39 animal profilins and 33 mammalian parvalbumins. The held-out allergens, as well as the presumed non-allergen proteins are publicly available on http://www.slv.se/templates/SLV_Page.aspx?id=14772.

RESULTS

Parameter selection

As detailed in the Materials and Methods, a holdout dataset was selected from the total allergen set to enable unbiased performance evaluation of the designed DFLAP system, whereas the remaining part was used for parameter selection in a 3-fold CV loop. In brief, eight values of the minimal peptide length $l_{\min}$, four retention levels (degree of filtration, the percentage of discarded allergen peptides), five values of the number of alignment scores n and five values of the SVM cost parameter C were tested in the CV loop. Thus, in total 800 ($8 \times 4 \times 5 \times 5$) different parameter settings were evaluated in the 3-fold CV.

The CV false alarm estimates were uniformly low for all different parameter settings (data not shown) suggesting that the general specificity of the DFLAP is quite robust. Therefore, parameter selection regarding specificity was focussed on returning parameter settings corresponding to few false alarms, also within particularly challenging protein families. Among tropomyosins, those of invertebrates are typically major allergens, whereas no vertebrate equivalent has hitherto been associated with hypersensitivity reactions (43). Nonetheless, this family shows high sequence similarity across several phylogenetic kingdoms. Consequently, the tropomyosins pose a serious challenge to bioinformatics protein allergen detection systems. Hence, we chose to obtain measures of specificity by taking advantage of the allergenicity dichotomy of tropomyosins. To obtain high performance on examples from this family, parameter settings were only considered useful if they resulted in false alarms CV estimates <10%, as regards to the five vertebrate tropomyosins placed in the non-allergen test set. Notably, this constraint allows only 1 of the 15 (five in each of three CV iterations) vertebrate tropomyosins to be erroneously assigned as an allergen. Amongst settings fulfilling this tropomyosin false alarm criterion, the setting yielding the highest CV estimate of the allergen detection rate was selected for the final detector design.

Frequencies of the four different parameters, occurring in the 80 best-ranked parameter settings that yielded the best detections at tropomyosin false alarms levels <10% are shown in Table 2 (column a). In order to identify parameters with the strongest impact on detection, the occurrence frequencies of the 80 best parameter settings—regardless of tropomyosins false alarms level—are also listed in Table 2 (column b). Analogously, to reveal the parameters important for correct assignment of tropomyosins, the occurrence frequencies among all the settings returning a tropomyosin false alarm level <10% were also calculated and are listed in Table 2 (column c).

Results in Table 2 clearly indicate that short peptides (low values of $l_{\min}$), aligned against the non-allergen database, correlate with low DFLAP performance regarding discriminations among tropomyosins. Actually, only few detectors based on peptide lengths below 16 residues could support assignment of the tropomyosins as non-allergens (Table 2, columns a and c). Moreover, there seems to be a negative correlation between decreased peptide lengths and DFLAP's overall detection capacity (Table 2, column b). In Table 2 (columns a and b) there are also indications on low values of the SVM cost parameter C accompanying poor

Table 2. Occurrence frequencies among the different parameters for the best detectors found according to a 3-fold CV

Occurrence frequencies among the different parameters (%)	a	b	c
Cost parameter			
$C = 0{,}1$	0	0	24
$C = 1$	0	0	22
$C = 10$	25	10	20
$C = 100$*	41	46	18
$C = 1000$	34	44	16
Peptide length			
$l_{\min} = 10$	0	0	0
$l_{\min} = 12$	0	11	0
$l_{\min} = 14$	0	4	2
$l_{\min} = 16$	4	13	17
$l_{\min} = 18$	3	6	17
$l_{\min} = 20$	21	18	21
$l_{\min} = 22$*	31	25	21
$l_{\min} = 24$	41	24	22
Retention level (filtration degree)			
75% retention	1	0	26
65% retention	15	13	27
55% retention	36	25	26
45% retention*	48	63	21
Number of matches			
$n = 1$	19	21	20
$n = 2$	23	20	20
$n = 3$	19	25	20
$n = 4$*	20	16	20
$n = 5$	20	18	20

*Preferred parameter setting in the finally selected detector.
(a) Summary of parameter settings returning the 80 highest detection rates, while at the same time showing tropomyosin false alarm estimates below 10%; (b) Summary of parameter settings producing the 80 highest detection levels, regardless of associated tropomyosin false alarm estimates; (c) Summary of parameter settings providing tropomyosin false alarm estimates below 10%, regardless of the associated detection performances.

detection estimates, whereas C does not seem to notably influence the assignment of vertebrate tropomyosins correctly as non-allergens (column c).

Another indication deduced from Table 2 is that a modest retention level is critical for obtaining good detection performance. For example, among the top (highest detection rates regardless of false alarm rate) 80 parameter settings, only 13% had retention level of 65% or higher (column b). The indication on a preference for low retention levels, with regard to successful detection, became strengthened by findings with settings that restricted the false alarm to low rates: The same degree of retention (65% or more) held for only 16% of the 80 best settings constrained to return low false alarm rate (Table 2, column a). On the opposite side of the performance spectrum, i.e. among the parameter settings providing tropomyosin false alarm estimates <10% regardless of the detection performance (column c), high retention levels seems to be preferred.

Finally, Table 2 also shows that the number n of matches (alignment scores) does not appear to markedly influence either the detection rate or the tropomyosins false alarms. Still, the optimum overall setting (i.e. the final detector design) includes the four best matches, indicating a predilection for several alignment scores. While parameter occurrence frequency is important to revealing the impact of the various parameters on performance, in practice only the overall top-ranked parameter setting was selected for final evaluation. Parameter values providing the highest detection estimate, on the provision that tropomyosin false alarm rates are <10% as assessed by a 3-fold CV, are as follows: $l_{\min} = 22$, FLAP threshold $= 48$ (corresponding to a retention level at 45%), $n = 4$ and $C = 100$.

Filtration and subsequent concatenation (when applicable) of all 762 allergens, applying the optimal values $l_{\min} = 22$ and a FLAP threshold $= 48$, resulted in 7196 FLAPs, equivalent to an average of almost 10 FLAPs for each allergen. The length distribution of the resulting FLAPs is depicted in Figure 2. The longest FLAPs encompass ~100 residues but only ~20% of them hold >40 residues. Roughly 50% of the FLAPs lie within the range of 22–28 residues. Notably, the shortest possible FLAP consists of a single, non-concatenated peptide of length $l_{\min} = 22$.

Performance comparison of DFLAP to other methods

As already presented, DFLAP performance was assessed in three separate test procedures: a holdout test of 262 allergens for sensitivity, a holdout test of 65 allergens and 193 non-allergens for intra-family discrimination ability and, finally, allergen detection in the entire Swiss-Prot database for specificity. For comparison purposes, these three experiments also served to evaluate the ILSI/IFBC, FAO/WHO and DASARP bioinformatics test schemes (see Table 1 and Materials and Methods for more details.)

Sensitivity for DFLAP and the three other tested procedures, indicating allergen detection, are presented as BC intervals for the different levels of sequence homology (maximal sequence similarity equal to 100, 95, 90, 85 and 80%, respectively) allowed in the test set (Figure 3). In addition, an overall measure of specificity was derived from tests based on the entire Swiss-Prot database as a query set (Table 3). With regard to the former benchmark type, the bioinformatics part of allergen identification, as prescribed by the FAO/WHO guidelines, significantly outperformed the methods included here. As evident from data presented in Table 3, however, the FAO/WHO approach is practically useless, owing to its high false alarms rate. Whereas the detection intervals of DFLAP seem slightly lower relative to that of ILSI/IFBC and higher to that of DASARP, the difference is quite small. Notably, though, the ILSI/IFBC counterpart features appreciably higher false alarm rates relative to that of DFLAP (Table 3).

In fact, DFLAP returned a remarkably small proportion of amino acid sequences assigned as allergens, outperforming corresponding readouts of the ILSI/IFBC- or FAO/WHO-proposed testing methods by a factor of at least 4, indicating a very low false-alarm level. The DFLAP benchmark result on specificity also outperformed that of DASARP (Table 3). Clearly, DFLAP is the most specific classifier, as demonstrated by an unmatched low allergen-assignment rate in the Swiss-Prot database.

DFLAP specificity and that of the three other methods, referred to above, were evaluated in an additional assessment procedure. This test involved challenge to allergens and presumable non-allergens from three distinct protein families. All methods were able to correctly assign the held-out allergens (data not shown), but as demonstrated for tropomyosins

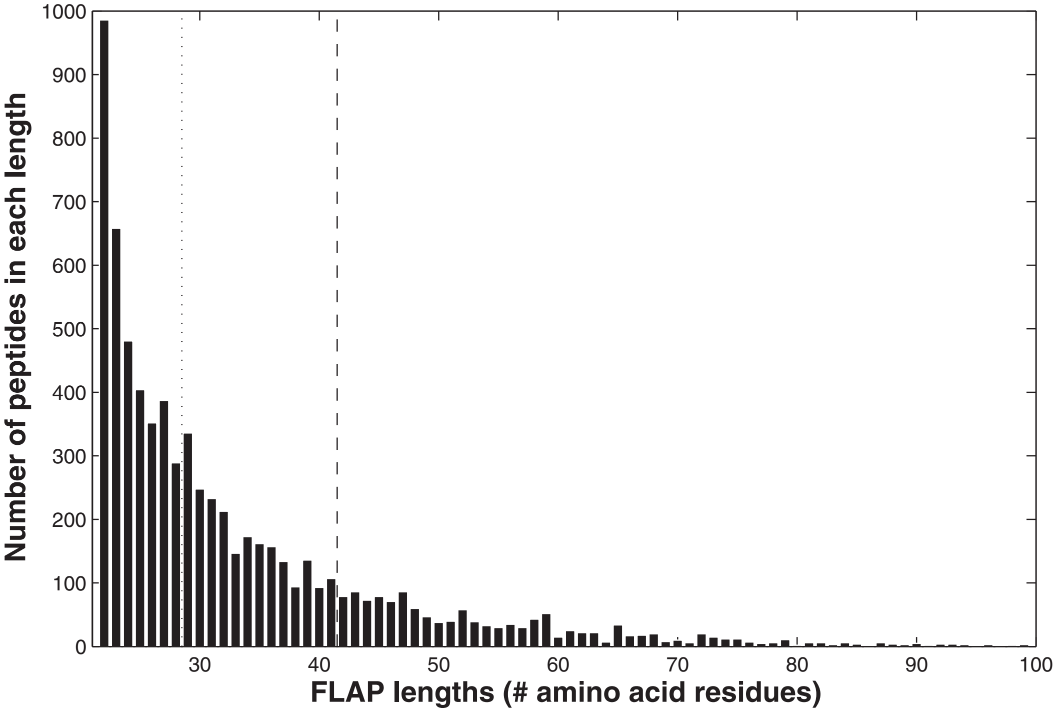

Figure 2. Length distribution of the final FLAP set based on 762 allergens (minimal peptide length, $l_{\min} = 22$ and FLAP threshold = 48). Of all FLAPs 50% are of length 28 or shorter (left part of dotted line) and 80% of length 41 or shorter (left part of dashed line).

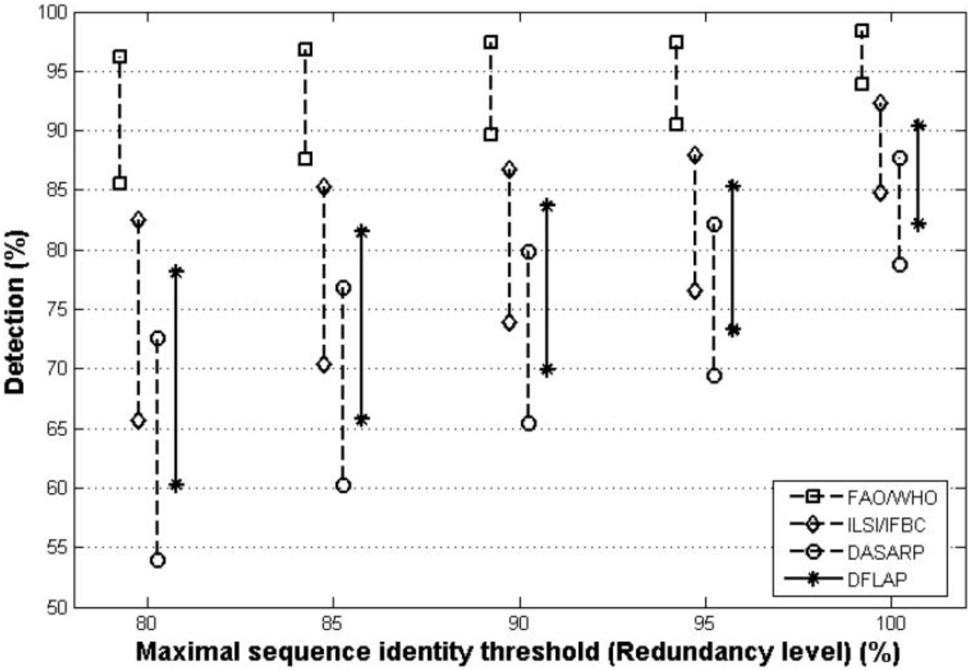

Figure 3. BC intervals (95%) of the unknown detection performance of the four tested methods using different levels of maximal sequence identity between training and test set. Clearly, the detection performance of the FAO/WHO method is much better than the other three but as shown in Figure 4, the corresponding false alarm rates make this approach useless. The DFLAP parameter setting was $l_{\min} = 22$, FLAP threshold = 48, $n = 4$ and $C = 100$.

and parvalbumins DFLAP was the only method to predict non-allergen protein family members outstandingly well (Figure 4). Also for profilins DFLAP performed reasonably well. In contrast, none of the comparator methods for allergenicity detection, except for ILSI/IFBC and DASARP in the case of profilins only, managed to qualify decently in this test (Figure 4). Thus, DFLAP represents a significant advancement as being the first *in silico* detector reported capable of accurate discrimination between allergens and non-allergens within several protein families.

DISCUSSION

Protein structural similarity (and difference) is an established fundament to understanding sensitization and cross-reactivity in allergy. In recent years, comparative scrutiny of plant food allergens has indicated that a predominant part of them fall within relatively few protein superfamilies (12,14). A highly biased distribution of plant food allergens across the protein structure universe is reported: about two-thirds of such molecules were found to occur in only four protein families (13). Analogously, pollen allergens are also confined to a small fraction of all known plant protein families (44). With a view to these findings a decent overall performance of allergen detectors is not unexpected for methods founded solely on similarity search over rather extended amino acid sequence segments, as typified by the FAO/WHO alignment procedure. Although many major allergens, such as the birch pollen allergen Bet v 1 showing high amino acid sequence similarity to some fruit proteins (e.g. Mal d 1 in apple and Api g 1 in celery) associated with IgE cross-reactive properties, there is, however, no universal correlation between sequence similarity of a protein to an allergen and its ability to trigger hypersensitivity type I-responses in atopic individuals. Hence, a major predicament in detection of allergenic potential in amino acid sequences is that many members of protein families known to hold a large proportion of allergens appear innocuous. This is particularly apparent for certain protein families holding members of high sequence similarity, but lacking allergenicity conservation, the tropomyosins being key examples thereof. Conversely, many members of the cupin superfamily display appreciable conservation of allergy but are poorly similar at the sequence level (12,45). In conclusion, algorithms developed to solely recognize common inter-allergen motifs will, in many instances, target protein motifs of little or no relevance to allergenicity. To confront constraints inherently connected with motif identification, regardless of scan window-size and approaches for recognition, within allergens only we have developed a profoundly different course focused on dissimilarities between allergens and non-offending proteins. We have earlier reported a prediction system founded on this general concept (35), but as clearly shown by the performance results, its full potential requires the several major novel features presented here.

Table 3. Estimated fractions of allergens in the Swiss-Prot database

Method	Swiss-Prot (1 64 970 samples) (%)
FAO/WHO	75.4
ILSI/IFBC	6.2
DASARP	3.1
DFLAP*	1.5

*The parameter setting was $l_{\min} = 22$, FLAP threshold = 48, $n = 4$ and $C = 100$.

To accomplish a selective enrichment of motifs, in the context of protein primary structure, that entail to allergen property the human proteome served as a vastly predominant source of amino acid sequences to help constructing the non-allergen database. Except for certain autoimmune disorders, typified by multiple sclerosis and systemic lupus

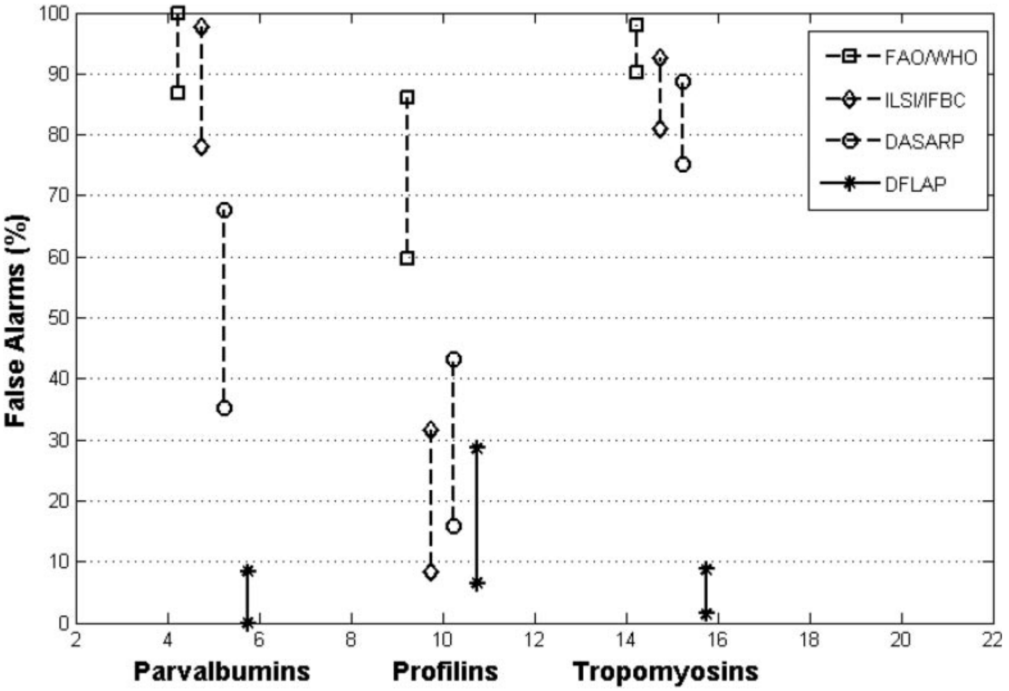

Figure 4. BC intervals (95%) of the false alarms for the four tested methods using (presumed) non-allergens belonging to three different protein families. Clearly, DFLAP is the only method that is able to discriminate successfully between allergens and non-allergens within the same protein family. The DFLAP parameter setting was $l_{\min} = 22$, FLAP threshold = 48, $n = 4$ and $C = 100$.

erythematosis, the immune system is generally tolerant to endogenous proteins. Therefore, we believe the non-allergen dataset being a conceptually apt target for removal of non-specific motifs in allergen amino acid sequences, as accomplished by the CPFA procedure. The principle suggesting that recognized epitopes generally share low similarity to the host's proteome has been applied in other areas as well. For example, Dummer *et al.* (46) has reported a computational scanning (using a non-self discrimination principle) of peptides derived from a melanoma antigen with the purpose of identifying epitopes. Certain allergens are highly dissimilar from any part of the human proteome. Consequently, the aforementioned CPFA step will not be able to generate any appreciable peptide reduction of those amino acid sequences. To address this problem we have included amino acid sequences from three additional species within two separate kingdoms, thereby providing a highly diverse set of complimentary sequences. Notably, any retrieved entity tagged with allergy, allergy-related features and proteins highly similar to them, was discarded prior to deposit in the non-allergen database. The additional sources are *A.fumigatus* (fungus), *C.albicans* (fungus) and *D.pteronyssinus* (dust mite). Their established usage in skin prick tests as whole extracts provides the rational for their inclusion in the non-allergen database after removal of reported allergens. Still, these proteomes are at somewhat higher risk of harboring non-documented immunogens, relative to the human counterpart, since only cross-reaction in already sensitized individuals can be clinically manifested in skin prick tests. We believe, however, the risk of contamination of the non-allergen database, as a consequence of unknown allergens/immunogens in these proteomes, being marginal. Detection rates, similar to those of ILSI/IFBC, support this conjecture (Figure 3). Although sequences from the three additional proteomes represent a very small fraction relative to the human counterpart their presence markedly affected the FLAP extraction procedure. For example, the filtration degree (percentage of peptides denied to entering the FLAP set) of fungi allergens increased with 9% when these sequences were included, as compared to a set based on the human proteome only (data not shown).

Performance: DFLAP compared with other bioinformatics methods

DFLAP's performance, in terms of sensitivity (detection of allergens), was evaluated with holdout validation as well as compared with results obtained by bioinformatics test for allergenicity according to guidelines proposed by FAO/WHO and ILSI/IFBC as well as to our recently reported DASARP algorithm. While the point estimates of DFLAP detection performances are in between those of ILSI/IFBC and DASARP (20,35) the BC intervals are almost totally overlapping (Figure 3). The BC intervals of FAO/WHO detections are, however, not overlapping with any of the other two procedures of comparison. The latter procedure has, however, attracted much criticism in recent years, from us and other researches, for being practically useless in testing for potential allergenicity (27,30,32,35,47), since too many false alarms are found with this method. This conclusion is further supported by findings in this study, revealing that >75% of all proteins occurring in Swiss-Prot were assigned as potential allergens by the FAO/WHO test procedure (Table 3). Regarding the relative assignment of allergens in Swiss-Prot, DFLAP showed a markedly lower estimate than any other method included in the comparative test. The DFLAP estimate (1.49%) is in fact the lowest number yet reported on this sort of performance test, using Swiss-Prot as an interrogator (32,33).

As outlined above, correct assignment of allergens within protein families poses an important and difficult challenge to detectors of potential allergenicity. It is generally accepted that tropomyosins from invertebrates, such as mites, shellfish and cockroaches, have the ability to elicit allergic reactions, whereas those originating from vertebrates are devoid of this characteristic (43). Actually, a recent report describes reduced IgE binding of the major shrimp allergen Pen a 1 (tropomyosin) in parallel with the gradual conversion to vertebrate tropomyosins by targeted substitution of key amino acid positions (48). In tests for specificity the DFLAP algorithm clearly outperformed the bioinformatics methods used here for comparison. Notably, DFLAP correctly assigned the 121 vertebrate tropomyosins as non-allergenic, whereas the other procedures produced false alarm rates of at least ~80% (Figure 4). In fact, DFLAP is the first prediction method proven to be able to classify vertebrate tropomyosins as non-allergens. In the motif-based prediction algorithm proposed by Li *et al.* (33), for example, it is stated that the motifs generated from allergenic tropomyosins are specific to family itself rather than the allergen counterpart. The mammalian parvalbumins and animal profilins included in our tests are not equally well documented on absence of hypersensitivity reactions in humans, but there is, however, reason to assign them as highly presumptive non-allergens. As far as we know there are no recognized allergens in any of these two data subsets. This indicates that they are suitable for testing the specificity of allergen detectors. In the case of mammalian parvalbumins almost all the 33 samples were classified as allergens, according to either of the FAO/WHO or ILSI/IFBC *in silico* protocol, whereas DASARP showed a false alarm rate near 50% (Figure 4). In contrast, DFLAP did not assign any of these sequences erroneously. Regarding the 39 animal profilins, only the FAO/WHO procedure assigned the vast majority of these presumed non-allergens as potential allergens, whereas the other three algorithms had overlapping BC intervals ranging from 5 to 45%. In brief, DFLAP features roughly the same sensitivity as that of either the ILSI/IFBC or DASARP but shows far better specificity, which is supported by the low estimate of the allergen frequency in Swiss-Prot. In addition, DFLAP can distinguish non-allergenic members from protein families known to hold allergens.

Notably, the method using either 35% sequence identity or an identical peptide match of eight contiguous amino acids (in this work referred to as ILSI/IFBC) as criteria for allergenicity, shows better overall performance than that based on a similar alignment procedure but an identical match motif of six amino acids, as recommended by FAO/WHO. Although the latter shows better sensitivity, it is has an unrealistic high detection rate (75%) of potential allergens in the Swiss-Prot database. Even though the DFLAP algorithm presented here and DASARP are founded on motif generation by comparison peptides of allergens to non-allergens rather than to other allergens, it clearly outperforms the latter procedure in all of the performed tests. We believe the improvements being partly owing to the introduction of a supervised learning machine, but the major advancement is likely to stem from the use of a more flexible peptide set, as compared to DASARP, regarding both peptide length and number of peptides per allergen. Future work is required to evaluate the relative influences of the different factors.

Addressing the issue of homology bias

Dedicated and publicly available repositories of protein allergens have proven indispensable for the development of computational methods for detecting allergen potential (49). In this work, data was mined from several such repositories and, subsequently to additional scrutiny, deposited in our in-house catalogue encompassing 762 protein allergen amino acid sequences. An appreciable part of the publicly listed allergens, though, occur as isoallergens, i.e. isoforms of the same allergenic protein. Relatively few allergens associate with many reported iso-forms, a considerably larger number have a few variants, whereas the majority of allergens occur as non-redundant forms. This representation may not, though, reflect the actual occurrence of allergen isoforms, because certain widely known sources of allergy are likely to have spewed targeted investigations on their respective proteins. In our allergen sequence archive there are, for example, more than 40 reported variants of the Bet v 1 birch pollen allergen; these variants may spread randomly into the design and validation sets. This sort of similarity also occurs across species boundaries. It is thus evident that this situation may substantially influence estimates of allergen detector performance. For example, an allergen amino acid sequence used for detector design, in conjunction with another isoform occurring in a performance evaluation set of examples, will inevitably facilitate the accurate identification of the latter. Thus, if redundant datasets are used in the design and validation of a detector, there is ample risk of obtaining overly optimistic performance estimates. This potential bias is well-known in most protein function prediction fields and removal of redundancy in datasets is commonly a standard operation prior to evaluation of performance. This issue has, however, not been extensively discussed in the literature in the context of allergen prediction and, as reported by Aalberse (50), many such studies have not taken this potential source of bias into consideration. In one of our earlier work (51), outlining a different allergen prediction approach (31) from that described here, a global sequence identity limit at 67% was employed to obtain non-redundant datasets. This limit can be regarded as an extension of one of the criteria for isoallergenicity, as proposed by WHO/IUIS Allergen Nomenclature Subcommittee (41).

In this work we have tackled the aforementioned iso-allergen dilemma by a step-wise reduction of sequence homology allowed. As illustrated in Figure 3, it is evident that detection rate estimations decline with decreasing redundancy in the dataset. Thus, current computational algorithms (including DFLAP), which are directed at searching for resemblance in sequence/structure to known allergens, are appropriate for predicting IgE-binding, but may less efficiently identify protein allergens that are highly dissimilar to those already known. It should be noted, however, that in contrast to many other protein function/structure problems, moderately high sequence identity to an allergen does not directly implicate allergenicity of the query protein. This is most conspicuous in the case of tropomyosins, showing high sequence similarity between allergens and non-allergens. With a too low sequence identity redundancy threshold, the performance validation may result in overly poor detection estimates. Accordingly, as it presently stands there is no general rational to help identifying an exact sequence identity redundancy threshold for validation of detectors of allergenic potential.

Parameter evaluation and the resulting FLAP set

We have recently described a source of bias that sometimes is overlooked in performance evaluation of computational algorithms for protein function/structure prediction (51). The problem arises when parameter tuning is not kept apart from performance evaluation in the validation process, which results in a biased performance estimate of the finally selected classifier. In the above mentioned study we proposed a double CV loop procedure wherein the internal one is used for parameter selection and the external for performance evaluation. The finally obtained external CV estimate(s) shows the robustness (or lack of robustness) of the total design procedure when tested on different datasets. While we in this work are interested in the accuracy of a single final detector, designed with settings as selected from the CV procedure, we have used holdout validation for performance evaluation rather than an external CV loop. This approach, which does not principally deviate from that referred to above (51), made it easier to compare the performance of DFLAP with those of some earlier reported bioinformatics algorithms.

As revealed by results listed in Table 2, the lowest limit for FLAP length lies in a narrow (four amino acids) range of ~20 amino acid residues. A detrimental impact of shorter lengths on accuracy regarding both classification of allergens (Table 2, column b) and non-allergenic tropomyosins (Table 2, column c) is evident. For a short minimal peptide length, e.g. five amino acids, almost all sliding-window penta-peptides will find a perfect match against the non-allergen database. If all peptides, either truly specific for allergy or truly not, would obtain roughly the same score it would be extremely difficult to identify peptides suitable for selection to the FLAP set. Some peptides truly specific for allergy would be rejected, whereas several of those unspecific in this context would be selected, resulting in a low-quality FLAP set. The low occurrence frequencies regarding shorter peptide lengths, listed in Table 2, supports this assumption. On the other hand, too extended lengths may imply high risk of relatively short segments, such as T-cell epitopes or linear IgE-epitopes, to escape detection. The reported length of such motifs range from 6 to ~20 amino acid residues, commonly 6–10 for linear segments of IgE-motifs and 9–20 for T-cell epitopes (7,52–56), the upper limit extending to the minimum FLAP lengths considered here. Potential future experiments to assess the potential role in allergenicity of specific FLAPs, could include *in vitro* testing, such as IgE-binding assays of FLAPs against sera from patients sensitized to the corresponding allergen, or comparison to experimentally verified epitopes. The work described in this article, however, is focused on the development and validation of a practical detector for protein allergens, rather than on structural modeling and/or functional identification of protein segments. Therefore, we have refrained from exhaustive testing of the nature of FLAPs, in terms of epitope or otherwise functional motifs. Moreover, we believe that a purpose oriented towards identification of allergen epitopes, in preference of high performance detection of allergens, should involve tuning of CPFA parameters based on different selection criteria, rather than those described and used in this work.

The alignment scores, derived from comparison to the non-allergen database and which are to be balanced to the FLAP threshold, are length-dependent. Hence, different FLAP thresholds were needed to compensate for peptide length variation and, accordingly, the threshold was adjusted to confine the proportion of discarded overlapping sliding-window peptides (percent retention) to roughly 45, 55, 65, or 75%. Intuitively, extensive filtration should increase the risk of rejecting peptides important for allergy. As demonstrated by comprehensive testing of parameters dictating high/low detector performance, a high filtration degree did not comply with good accuracy (Table 2, columns a and b). On the other hand, relaxed filtration would also make it harder to assign non-allergens correctly, since it imparts higher risk of incorporating peptides unrelated to allergy into the FLAP set. The occurrence frequency of the lowest retention level (45%) seems to be slightly lower than the others (Table 2, column c), although the differences are rather small (21% as compared with 26 and 27% frequency). The lower relative occurrence (48% as compared with 63% frequency) at a retention level of 45% amongst the 80 top-ranked settings fulfilling the tropomyosins false alarm level criterion (Table 2, column a), compared with the best 80 settings regardless of this criterion (Table 2, column b), further support association between low filtration degree and poor DFLAP ability to discriminate among allergen/non-allergen tropomyosins.

The parameter selection process revealed an appreciable influence of the SVM cost (or regularization) parameter C on detection performance (Table 2, columns a and b). However, the level of false alarm among non-allergenic vertebrate tropomyosins appeared to be quite insensitive to changes of C (column c). Whereas the sampling of values for C were rather sparse (step-wise increments by a factor of 10), a more refined search was employed around the value of C occurring in the best parameter setting, while other optimized parameters were kept constant. We found, however, no divergence at all between results from the SVMs based on different values of C (data not shown). Whether we will get relatively few or many training errors for a given value of the parameter C cannot be predicted; it will depend on the particular dataset used. Moreover, this kind of SVM design only involves a penalty associated with misclassification regardless of whether it represents escaped detection or false alarm. Thus, the SVM design procedure itself is not explicitly designed to minimize the number of misclassifications or to balance the two plausible types of misclassifications. Therefore, the fact that detection performance but not false alarm rate seems to be sensitive to the parameter C in Table 3 indicates that the actual SVM learning algorithm used is biased towards minimization of missed detections, rather than on minimization of false alarm in this particular application of allergen detection. Adapting the SVM learning algorithm to high susceptibility to false alarm rate is certainly interesting, but not further elaborated on in this article.

Seemingly, (Table 2) the number of best matches (alignment scores), n, of a query sequence against the FLAP is a parameter with low influence on detection and false alarm. While allergens have multiple epitopes, the appearance of roughly equal relative frequencies of the single best match ($n = 1$) and other settings ($n = 2, 3, 4, 5$) for optimal classification may appear unexpected. A minimal-length FLAP, though, holds 22 amino acid residues and, as illustrated in Figure 2, 50% of all FLAPs encompass 29 residues or more, i.e. a single FLAP could principally embrace multiple epitopes. In addition, we do not claim the FLAPs to be truly defined epitopes; they may also hold other kinds of structural motifs indirectly important to allergenicity. The use of a supervised machine learning algorithm may seem to be superfluous in the special case when only one score value ($n = 1$) against the FLAP set is used to represent a protein's feature vector, since the resulting decision surface will only correspond to a simple point threshold. Nonetheless, the ultimately selected—thereby highest ranked—detector is founded on four score values. Moreover, consistency in the design procedure during parameter evaluation is maintained by taking advantage of the SVM for all values of n.

CONCLUSIONS

A large body of computational methods for *in silico* detection of allergens has been reported. Until now, however, none of them have been successful with respect to overall specificity as well as discrimination between allergens and non-allergens in particularly challenging homologous protein families, such as the tropomyosins. For the first time, this work shows that it is possible to design new computational detectors that successfully confront both these problems. In particular, one such detector, designated DFLAP, has been presented that extracts and employs allergen representative peptides with variable lengths for sequence feature extraction and uses modern machine learning techniques for detector design.

The significant improvements of DFLAP may be illustrated by a careful methodological comparison with DASARP, a detector that has been reevaluated in this article and relies on extraction and employment of fixed length peptides for feature extraction and a simple decision procedure. The results of this comparison indicate that the significant improvements rely on a combination of biologically more relevant features owing to flexible peptides and improved fine-tuning of the computational decision process, as accomplished by the modern machine learning employed.

ACKNOWLEDGEMENTS

The authors thank Ms Maj Olausson and Mrs Merethe Andersen for their expert assistance on computer graphics.

The Cancer and Allergy Fund (Stockholm), Carl Tryggers Stiftelse (Stockholm), the Göran Gustafsson foundation (Stockholm), and the Faculty of Science and Technology (Uppsala University), are acknowledged for financial support of this work.

Conflict of interest statement. None declared.

REFERENCES

1. Burney,P., Malmberg,E., Chinn,S., Jarvis,D., Luczynska,C. and Lai,E. (1997) The distribution of total and specific serum IgE in the European Community Respiratory Health Survey. *J. Allergy Clin. Immunol.*, **99**, 314–322.
2. Broadfield,E., McKeever,T.M., Scrivener,S., Venn,A., Lewis,S.A. and Britton,J. (2002) Increase in the prevalence of allergen skin sensitization in successive birth cohorts. *J. Allergy Clin. Immunol.*, **109**, 969–974.
3. Platts-Mills,T.A. (2001) The role of immunoglobulin E in allergy and asthma. *Am. J. Respir. Crit. Care. Med.*, **164**, S1–5.
4. Miescher,S.M. and Vogel,M. (2002) Molecular aspects of allergy. *Mol. Aspects Med.*, **23**, 413–462.
5. Johansson,S.G., Hourihane,J.O., Bousquet,J., Bruijnzeel-Koomen,C., Dreborg,S., Haahtela,T., Kowalski,M.L., Mygind,N., Ring,J., van Cauwenberge,P. *et al.* (2001) A revised nomenclature for allergy. An EAACI position statement from the EAACI nomenclature task force. *Allergy*, **56**, 813–824.
6. Kay,A.B. (2001) Allergy and allergic diseases. Second of two parts. *N. Engl. J. Med.*, **344**, 109–113.
7. Aalberse,R.C., Akkerdaas,J. and van Ree,R. (2001) Cross-reactivity of IgE antibodies to allergens. *Allergy*, **56**, 478–490.
8. Weber,R.W. (2001) Cross-reactivity of plant and animal allergens. *Clin. Rev. Allergy Immunol.*, **21**, 153–202.
9. Vieths,S., Scheurer,S. and Ballmer-Weber,B. (2002) Current understanding of cross-reactivity of food allergens and pollen. *Ann. NY Acad. Sci.*, **964**, 47–68.
10. Yagami,T. (2002) Allergies to cross-reactive plant proteins. Latex-fruit syndrome is comparable with pollen-food Allergy syndrome. *Int. Arch. Allergy Immunol.*, **128**, 271–279.
11. Breiteneder,H. and Ebner,C. (2001) Atopic allergens of plant foods. *Curr. Opin. Allergy Clin. Immunol.*, **1**, 261–267.
12. Breiteneder,H. and Radauer,C. (2004) A classification of plant food allergens. *J. Allergy Clin. Immunol.*, **113**, 821–830.
13. Jenkins,J.A., Griffiths-Jones,S., Shewry,P.R., Breiteneder,H. and Mills,E.N. (2005) Structural relatedness of plant food allergens with specific reference to cross-reactive allergens: an *in silico* analysis. *J. Allergy Clin. Immunol.*, **115**, 163–170.
14. Mills,E.N., Jenkins,J.A., Alcocer,M.J. and 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.
15. van der Zee,J.S., de Groot,H., van Swieten,P., Jansen,H.M. and Aalberse,R.C. (1988) Discrepancies between the skin test and IgE antibody assays: study of histamine release, complement activation *in vitro*, and occurrence of allergen-specific IgG. *J. Allergy Clin. Immunol.*, **82**, 270–281.
16. Kimber,I., Dearman,R.J., Penninks,A.H., Knippels,L.M., Buchanan,R.B., Hammerberg,B., Jackson,H.A. and Helm,R.M. (2003) Assessment of protein allergenicity on the basis of immune reactivity: animal models. *Environ. Health Perspect.*, **111**, 1125–1130.
17. Hamilton,R.G. and Adkinson,N.F.Jr (2004) *In vitro* assays for the diagnosis of IgE-mediated disorders. *J. Allergy Clin. Immunol.*, **114**, 213–225.
18. Konig,A., Cockburn,A., Crevel,R.W., Debruyne,E., Grafstroem,R., Hammerling,U., Kimber,I., Knudsen,I., Kuiper,H.A., Peijnenburg,A.A. *et al.* (2004) Assessment of the safety of foods derived from genetically modified (GM) crops. *Food Chem. Toxicol.*, **42**, 1047–1088.
19. Goodman,R.E., Hefle,S.L., Taylor,S.L. and van Ree,R. (2005) Assessing genetically modified crops to minimize the risk of increased food allergy: a review. *Int. Arch. Allergy Immunol.*, **137**, 153–166.
20. Metcalfe,D.D., Astwood,J.D., Townsend,R., Sampson,H.A., Taylor,S.L. and Fuchs,R.L. (1996) Assessment of the allergenic potential of foods derived from genetically engineered crop plants. *Crit. Rev. Food Sci. Nutr.*, **36** (Suppl), S165–S186.
21. FAO/WHO. (2001) *Joint FAO/WHO Expert Consultation on Allergenicity of Foods Derived from Biotechnology,* Rome, Italy.
22. Codex Alimentarius Comission. (2004) *Foods Derived from Biotechnology. Joint FAO/WHO Food Standards Programme,* Food and Agriculture Organisation of the UN WHO, Rome, Italy.
23. Ivanciuc,O., Schein,C.H. and Braun,W. (2003) SDAP: database and computational tools for allergenic proteins. *Nucleic Acids Res.*, **31**, 359–362.
24. Fiers,M.W., Kleter,G.A., Nijland,H., Peijnenburg,A.A., Nap,J.P. and van Ham,R.C. (2004) Allermatch, a webtool for the prediction of potential allergenicity according to current FAO/WHO Codex alimentarius guidelines. *BMC Bioinformatics*, **5**, 133.
25. Nakamura,R., Teshima,R., Takagi,K. and Sawada,J. (2005) [Development of Allergen Database for Food Safety (ADFS): an integrated database to search allergens and predict allergenicity]. *Kokuritsu Iyakuhin Shokuhin Eisei Kenkyusho Hokoku*, 32–36.
26. Gendel,S.M. (1998) 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.
27. Gendel,S.M. (2002) Sequence analysis for assessing potential allergenicity. *Ann. NY Acad. Sci.*, **964**, 87–98.
28. Zorzet,A., Gustafsson,M. and Hammerling,U. (2002) Prediction of food protein allergenicity: a bioinformatic learning systems approach. *In. Silico Biol.*, **2**, 525–534.
29. Hileman,R.E., Silvanovich,A., Goodman,R.E., Rice,E.A., Holleschak,G., Astwood,J.D. and Hefle,S.L. (2002) Bioinformatic methods for allergenicity assessment using a comprehensive allergen database. *Int. Arch. Allergy Immunol.*, **128**, 280–291.
30. Kleter,G.A. and Peijnenburg,A.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.
31. Soeria-Atmadja,D., Zorzet,A., Gustafsson,M.G. and Hammerling,U. (2004) Statistical evaluation of local alignment features predicting allergenicity using supervised classification algorithms. *Int. Arch. Allergy Immunol.*, **133**, 101–112.
32. Stadler,M.B. and Stadler,B.M. (2003) Allergenicity prediction by protein sequence. *FASEB J.*, **17**, 1141–1143.
33. Li,K.B., Issac,P. and Krishnan,A. (2004) Predicting allergenic proteins using wavelet transform. *Bioinformatics*, **20**, 2572–2578.
34. Ivanciuc,O., Schein,C.H. and Braun,W. (2002) Data mining of sequences and 3D structures of allergenic proteins. *Bioinformatics*, **18**, 1358–1364.
35. Bjorklund,A.K., Soeria-Atmadja,D., Zorzet,A., Hammerling,U. and Gustafsson,M.G. (2005) Supervised identification of allergen-representative peptides for *in silico* detection of potentially allergenic proteins. *Bioinformatics*, **21**, 39–50.
36. Boeckmann,B., Bairoch,A., Apweiler,R., Blatter,M.C., Estreicher,A., Gasteiger,E., Martin,M.J., Michoud,K., O'Donovan,C., Phan,I. *et al.* (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.*, **31**, 365–370.
37. Webb,A. (2002) *Statistical Pattern Recognition, 2nd edn.* Wiley, Chicester.
38. Hastie,T., Tibshirani,R. and Friedman,J. (2001) *The Elements of Statistical Learning: Data Mining, Inference and Prediction.* Springer-Verlag, New York.
39. Pearson,W.R. (2000) Flexible sequence similarity searching with the FASTA3 program package. *Methods Mol. Biol.*, **132**, 185–219.
40. Jaynes,E.T. (1976) Confidence Intervals vs Bayesian Intervals. In Harper,W.L. and Hooker,C.A. (eds), *Foundations of Probability Theory, Statistical Inference, and Statistical Theories of Science.* D.Reidel, Dordrecht, pp. 175–257.
41. IUIS/WHO. (1994) Allergen nomenclature. IUIS/WHO. Allergen Nomenclature Subcommittee. *Bull. World Health Organ.*, **72**, 797–806.
42. Gendel,S.M. (1998) Sequence databases for assessing the potential allergenicity of proteins used in transgenic foods. *Adv. Food Nutr. Res.*, **42**, 63–92.
43. Reese,G., Ayuso,R. and Lehrer,S.B. (1999) Tropomyosin: an invertebrate pan-allergen. *Int. Arch. Allergy Immunol.*, **119**, 247–258.

44. Radauer,C. and Breiteneder,H. (2006) Pollen allergens are restricted to few protein families and show distinct patterns of species distribution. *J. Allergy Clin. Immunol.*, **117**, 141–147.

45. Wensing,M., Knulst,A.C., Piersma,S., O'Kane,F., Knol,E.F. and Koppelman,S.J. (2003) Patients with anaphylaxis to pea can have peanut allergy caused by cross-reactive IgE to vicilin (Ara h 1). *J. Allergy Clin. Immunol.*, **111**, 420–424.

46. Dummer,R., Mittelman,A., Fanizzi,F.P., Lucchese,G., Willers,J. and Kanduc,D. (2004) Non-self-discrimination as a driving concept in the identification of an immunodominant HMW-MAA epitopic peptide sequence by autoantibodies from melanoma cancer patients. *Int. J. Cancer*, **111**, 720–726.

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

48. Reese,G., Viebranz,J., Leong-Kee,S.M., Plante,M., Lauer,I., Randow,S., Moncin,M.S., Ayuso,R., Lehrer,S.B. and Vieths,S. (2005) Reduced allergenic potency of VR9-1, a mutant of the major shrimp allergen Pen a 1 (tropomyosin). *J. Immunol.*, **175**, 8354–8364.

49. Brusic,V., Millot,M., Petrovsky,N., Gendel,S.M., Gigonzac,O. and Stelman,S.J. (2003) Allergen databases. *Allergy*, **58**, 1093–1100.

50. Aalberse,R.C. (2005) Assessment of sequence homology and cross-reactivity. *Toxicol. Appl. Pharmacol.*, **207**, 149–151.

51. Soeria-Atmadja,D., Wallman,M., Bjorklund,A.K., Isaksson,A., Hammerling,U. and Gustafsson,M.G. (2005) External cross-validation for unbiased evaluation of protein family detectors: Application to allergens. *Proteins*, **61**, 918–925.

52. Burks,A.W., Shin,D., Cockrell,G., Stanley,J.S., Helm,R.M. and Bannon,G.A. (1997) Mapping and mutational analysis of the IgE-binding epitopes on Ara h 1, a legume vicilin protein and a major allergen in peanut hypersensitivity. *Eur. J. Biochem.*, **245**, 334–339.

53. Khalil-Daher,I., Boisgerault,F., Feugeas,J.P., Tieng,V., Toubert,A. and Charron,D. (1998) Naturally processed peptides from HLA-DQ7 (alpha1*0501-beta1*0301): influence of both alpha and beta chain polymorphism in the HLA-DQ peptide binding specificity. *Eur. J. Immunol.*, **28**, 3840–3849.

54. Rudensky,A., Preston-Hurlburt,P., Hong,S.C., Barlow,A. and Janeway,C.A. Jr (1991) Sequence analysis of peptides bound to MHC class II molecules. *Nature*, **353**, 622–627.

55. Sant'Angelo,D.B., Robinson,E., Janeway,C.A. Jr and Denzin,L.K. (2002) Recognition of core and flanking amino acids of MHC class II-bound peptides by the T cell receptor. *Eur. J. Immunol.*, **32**, 2510–2520.

56. Stanley,J.S., King,N., Burks,A.W., Huang,S.K., Sampson,H., Cockrell,G., Helm,R.M., West,C.M. and Bannon,G.A. (1997) Identification and mutational analysis of the immunodominant IgE binding epitopes of the major peanut allergen Ara h 2. *Arch. Biochem. Biophys.*, **342**, 244–253.