



Supervised identification of allergen-representative peptides for in silico detection of potentially allergenic proteins

Åsa K. Björklund^{1,†}, Daniel Soeria-Atmadja¹, Anna Zorzet¹, Ulf Hammerling^{1,*} and Mats G. Gustafsson²

¹Division of Toxicology, National Food Administration, P.O. Box 622, SE-751 26 Uppsala, Sweden and ²Department of Engineering Sciences, Uppsala University, P.O. Box 528, SE-751 20 Uppsala, Sweden

Received on December 4, 2003; revised on July 2, 2004; accepted on August 5, 2004
Advance Access publication August 19, 2004

ABSTRACT

Motivation: Identification of potentially allergenic proteins is needed for the safety assessment of genetically modified foods, certain pharmaceuticals and various other products on the consumer market. Current methods in bioinformatic allergology exploit common features among allergens for the detection of amino acid sequences of potentially allergenic proteins. Features for identification still unexplored include the motifs occurring commonly in allergens, but rarely in ordinary proteins. In this paper, we present an algorithm for the identification of such motifs with the purpose of biocomputational detection of amino acid sequences of potential allergens.

Results: Identification of allergen-representative peptides (ARPs) with low or no occurrence in proteins lacking allergenic properties is the essential component of our new method, designated DASARP (Detection based on Automated Selection of Allergen-Representative Peptide). This approach consistently outperforms the criterion based on identical peptide match for predicting allergenicity recommended by ILSI/IFBC and FAO/WHO and shows results comparable to the alignment-based criterion as outlined by FAO/WHO.

Availability: The detection software and the ARP set needed for the analysis of a query protein reported here are properties of the Swedish National Food Agency and are available upon request. The protein sequence sets used in this work are publicly available on http://www.slv.se/templatesSLV/SLV_Page___9343.asp. Allergenicity assessment for specific protein sequences of interest is also possible via ulfh@slv.se

Contact: ulfh@slv.se

INTRODUCTION

Atopic allergy and other forms of hypersensitivity affect up to 15–20% of the population in industrial nations. The estimated prevalence of food allergenicity among the general population within the European Union ranges from about 2.5 to 3.2% (Jansen *et al.*, 1994; Kanny *et al.*, 2001; SCP, 1998). Typical allergy (Type I hypersensitivity reaction) symptoms are rhinitis, asthma and atopic eczema, but more severe reactions such as acute and possibly fatal anaphylactic shock can also occur. Allergy is caused by adverse immune responses to otherwise innocuous proteins, the allergens. In atopic individuals, sensitizing T-cell epitopes can trigger a cascade of events leading to synthesis of allergen-specific immunoglobulin E (IgE) antibodies as well as other immunological reactions. The IgE antibodies bind the intruding allergen, or a structurally similar cross-reacting protein, leading to the release of mediators, which causes allergic reactions. Hence, T- and B- (IgE) epitopes are both relevant targets for models aimed at the detection of protein allergens. The former type is generally confined to continuous motifs of about 8–24 amino acid residues, whereas the latter may occur as scattered regions, which are brought together on the three-dimensional surface of the protein (Bredehorst and David, 2001).

Implementation of molecular genetics in the pharmaceutical industry, plant breeding corporations and among producers of household biochemicals has increased over the last decade resulting in an appreciable number of recombinant proteins on the market (Chuang *et al.*, 2002; Gualandi-Signorini and Giorgi, 2001; Heppenheimer, 2003; Lee and Sinko, 2000). Moreover, emerging bio-pharmaceuticals for oral administration and foods derived from transgenic animals result in human exposure to an even broader spectrum of such bio-molecules (Kuehn, 2003; Soltero and Ekwuribe, 2002). Accordingly, there is increasing need for safety assessment. In recent years, much attention has been paid to safety issues in the context of genetically modified (GM) food crops, where

*To whom correspondence should be addressed.

[†] Present address: Stockholm Bioinformatics Center, SCFAB, Stockholm University, SE-10691 Stockholm, Sweden

a key issue is the risk associated with the introduction of xenoproteins having the potential to induce allergic responses in consumers. In addition to laboratory experimentation and testing in clinical settings, current procedures for allergenicity assessment involve an introductory comparison of the novel protein's amino acid sequence with those of known allergens (FAO/WHO, 2001, 2003; Metcalfe *et al.*, 1996). An expert consultation on foods derived from biotechnology, convened by the FAO/WHO (Food and Agriculture Organization/World Health Organization), reported on an evaluation scheme for potential allergenicity in 2001 (FAO/WHO, 2001). In this scheme, the initial test procedure involves comparison of the novel protein's amino acid sequence with those of known allergens, specifying either a match of six consecutive amino acids or 35% homology over 80 amino acids as being indicative of potential allergenicity (FAO/WHO, 2001). Over the last few years, however, several reports have presented criticism against the first criterion of this procedure as being fairly non-specific (Goodman *et al.*, 2002; Hileman *et al.*, 2002; Kleter and Peijnenburg, 2002; Soeria-Atmadja *et al.*, 2004). Additionally, while the FAO/WHO report only refers to general protein databases, several specialized repositories of amino acid sequences associated with allergy reactions have become available publicly (Brusic *et al.*, 2003; FAO/WHO, 2001).

Owing to these problems, several alternative courses to the identification of allergy-indicative sequences similarity have been described recently. Gendel has proposed a strategy that involves an initial search for matches of identical amino acids, followed by inspection using either a biochemical or an evolutionary substitution matrix (Gendel, 1998b, 2002). Testing for an identical peptide match using a sliding window of either of 6, 7 or 8 amino acids as well as FASTA-based searches has been reported (Hileman *et al.*, 2002). Other directions involve primary sequence or structural comparisons with IgE epitope data from literature, either as a stand-alone method or in combination with searches for short identical stretches (Ivanciuc *et al.*, 2002, 2003; Kleter and Peijnenburg, 2002).

Moreover, identification of regions typical for allergenic proteins by using an iterative motif-finding approach was described recently (Stadler and Stadler, 2003). In conclusion, the literature on bioinformatics approaches to the detection of potential allergenicity is focused on identifying inter-allergen similarities rather than considering the potential of features that differ consistently between examples of allergens and non-allergens. Except for a limited number of food proteins (Hileman *et al.*, 2002), there are no reported results on the performance, in the context of allergen prediction, of 35% sequence identity over an 80 amino acid window.

Features that differentiate allergens from non-allergens are difficult to find by manual inspection of amino acid sequences. We have recently reported a bioinformatic learning systems approach where examples of both allergens and non-allergens from commonly eaten commodities were used to design detectors of potentially allergenic proteins (Zorzet

et al., 2002). As the first example of this approach, each allergen and non-allergen sequence was aligned to a set of pre-defined allergen prototype sequences yielding a characteristic pattern of alignment scores and alignment lengths. These characteristic patterns were then used as inputs to different forms of supervised classification algorithms. In a more recent study (Soeria-Atmadja *et al.*, 2004), we describe examples of markedly more refined systems, which employ a larger dataset, three different supervised classification algorithms, and various kinds of parameter settings for the alignment-based feature extraction.

Even though these already reported supervised detectors of potential allergenicity performed well, we saw room for improvement by implementing an alternative method. Detectors of allergenic potential based on alignment descriptors are sensitive to cross-reacting allergenic proteins due to structural similarities, but may still overlook other similarities important for allergic reactions. Accordingly, the selection of appropriate descriptors for accurate allergen prediction poses a tough challenge. We find approaches based on supervised peptide identification procedure, designated Automated selection of allergen-representative peptides (ASARP), to be a powerful alternative or complement to traditional alignment-based methods. Such approaches have strong algorithmic benefit since they can operate independently of pre-selected sets of allergen prototypes. Our earlier work shows that the selection of such a set is critical for the final performance. Moreover, this pre-selection reduces the number of remaining allergens available for statistical validation of the final detector of allergenic potential.

With this in mind, we have developed a supervised classification approach to *in silico* detection of potentially allergenic proteins, founded on the automated selection of allergen-representative peptides (ARPs) and released from pre-selection of prototypes. We first introduce the supervised algorithm ASARP for identification of ARPs and then demonstrate the potential of ARPs for allergen prediction. ASARP is based on two peptide repositories: one from sequences selected for being devoid of any connection to allergy, and one from a set of allergens retrieved from publicly available databases. Comparison of two such sets, and the subsequent selection of the allergen-representative peptides that are least similar to the non-allergens, are the key steps leading to the establishment of an ARP set. The novel detection approach introduced here, Detection based on ASARP (DASARP), employs the set of ARPs, thereby offering a unique detection method that specifically employs portions of the allergens that are either absent or occur rarely in non-allergens. The novel detector has been evaluated extensively using 10-fold cross-validation for a limited number of different parameter settings (additional tuning would increase the risk of overfitting). DASARP yields improved *in silico* detection of a protein's potential allergenicity compared with our previously reported methods as well as the identical peptide match procedure

recommended by the FAO/WHO and the ILSI/IFBC. The performance is comparable to the alignment-based method, as recommended by the FAO/WHO (2001, 2003) and could therefore be applied to the evaluation of potential allergenicity of GM food-encoded proteins, as a part of an integrated risk assessment procedure.

SYSTEMS AND METHODS

Establishment of in-house databases with allergenic and non-allergenic sequences

High-quality repositories of amino acid sequences of proteinaceous allergens and non-allergen counterparts provide a prerequisite for the establishment of ARPs. For this purpose, amino acid sequences of allergenic proteins were mined from the following six publicly available databases: Allergen nomenclature (King *et al.*, 1995), Farrp (Hileman *et al.*, 2002), The Allergen Database (<http://www.csl.gov.uk/allergen/>), The Allergen Sequence Database (Gendel, 1998a), The Protall Database (<http://www.ifrn.bbsrc.ac.uk/protall/>) and The Allergome Database (<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 such poor documentation were omitted from the allergen repository. Additionally, sequences occurring as fragments shorter than 100 amino acids were removed to reduce the risk of incorporating protein allergens without allergy-inducing (sensitizing or cross-reactive) regions.

To create useful ARPs, a large repository of amino acid sequences of non-allergenic proteins is required, since it ideally should contain all possible peptides occurring in any protein devoid of allergenic properties. A repository with broad representation of protein sequences, selected to be devoid of allergenicity, was accomplished by mining the Swall database (Boeckmann *et al.*, 2003) using the ExPaSy sequence retrieval system (Gasteiger *et al.*, 2003). The (presumably) non-allergenic sequences were excerpted from commonly consumed commodities, such as rice, apple, carrot, peach, cherry, apricot, spinach, tomato, salmon, cow's milk and hen's egg. These species were selected to minimize the risk of introducing unknown allergens in this set, based on the following assumption: while allergens occur in the species selected, they are thoroughly inspected for allergens, inferring that the remaining proteins from these commodities are at relatively low risk of being allergenic. All proteins from these species and available in Swall were selected, except for entries containing the text strings 'allergen', 'allergy' or 'atopy' as well as sequences being shorter than 50 amino acids. An additional inspection was done to confirm that no accordingly retrieved sequences appeared in any of the allergen databases mentioned above.

Automated selection of ARPs

A computer program for ASARP, based on a new algorithm, was created. The algorithm's function can be summarized as follows: given two repositories of protein sequences, one containing allergens and the second one containing non-allergens, perform the following steps: (1) segment all protein amino acid sequences (allergens and non-allergens) into peptides of a pre-defined length (l) by means of a sliding window. (2) For each peptide in the allergen repository, compute a similarity score based on alignment, which reflects the degree of similarity for that peptide relative to each peptide in the non-allergen database. (3) From these individual similarity scores computed for each allergen peptide, merge them into one or several global similarity scores, which reflect the similarity between the individual peptide and the whole set of non-allergen peptides, as in hierarchical clustering. (4) Based on these global similarity scores of each allergen peptide, create a set containing a predefined number (n_a) of ARPs originating from each allergen yielding in total N_a ($N_a = n_a \cdot$ number of allergens) allergen peptides. This procedure is depicted schematically in Figure 1.

More specifically, in Step (1) and subsequent to partition, peptides in the non-allergen repository occurring only once were removed. Our aim is to identify sub-sequences, which occur frequently in non-allergens, and to eliminate them from our set of allergen peptides. Hence, non-allergen peptides with only a single copy in the whole set (about 10^6 peptides) were assumed to be dispensable. In Step (2) the individual similarities were defined as: $\sum_{i=1}^l s(X_i Y_i)$, where $s(X_i Y_i)$ is the substitution score between amino acid i in peptide X and the amino acid i in peptide Y , according to the substitution matrix employed (Blosom80 or Pam30). This measure is related to non-gapped alignment scores. The individual scores were then merged into global similarity scores according to two different approaches. In the first, the similarity between a single allergen peptide and the set of non-allergen peptides was defined as the number of individual similarity scores found above a pre-defined threshold (T_A). Several values for this threshold T_A were tested and we found that variations in the threshold did not appreciably influence the results. In the data presented here a threshold (T_A) at 2, providing the best results, was used. Thus, only a single global feature, the number of 'hits' (n_h) above this threshold was computed. In the second approach, a feature list ($[S_{A1}, S_{A2}, \dots, S_{A10}]$) containing the 10 highest alignment scores was extracted for each peptide.

In Steps (3) and (4), several possible procedures for the final selection of the ARPs were tested. In this paper, evaluations of the three most promising ones are described: (i) select the peptides with the lowest value of n_h . This procedure is intuitive as it selects peptides that have none or very few (defined by the threshold T_A) matching non-allergen peptides. This measure of similarity is, however, not robust against errors in the manual labelling of the non-allergen peptides.

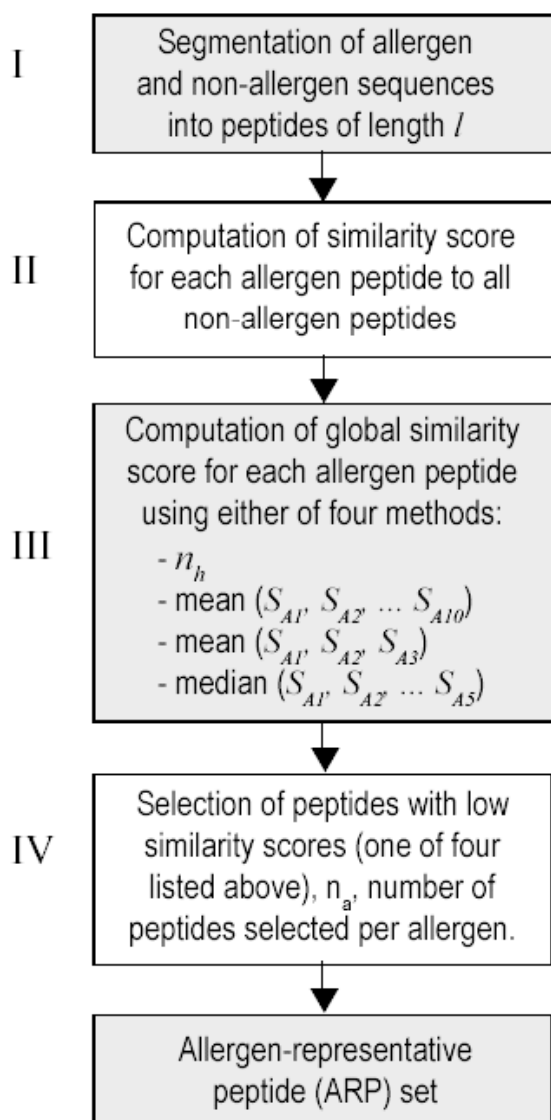


Fig. 1. A flowchart showing steps involved in ASARP. (I) First the two repositories of protein sequences (allergens and non-allergens) are segmented into peptides of length l by means of a sliding window. (II) A similarity score, based on alignment, reflecting the degree of similarity for each peptide in the allergen repository relative to all the non-allergen peptides, is calculated. (III) From these scores, one or several global scores that reflect the similarity between the individual peptide and the whole set of non-allergen peptides is calculated in four different manners. (IV) Based on the global similarity scores of each allergen peptide, a set containing a defined number (n_a) of ARPs originating from each allergen is created, yielding in total N_a allergen peptides. These ARPs are subsequently used for biocomputational detection of potential allergens.

(ii) Select the peptides with the smallest arithmetic mean value of the three (or ten) highest score values in the feature vector $[S_{A1}, S_{A2}, \dots, S_{A10}]$. This procedure is also intuitive as it accomplishes selection of peptides, which—also for the three

(or ten) best matching non-allergen peptides—yields very low average scores. In other words, a low similarity score for the best matches implies a low global similarity score with the set of non-allergen peptides. Although this approach is intuitive, the arithmetic mean is not entirely robust against errors in the labelling of non-allergen peptides either. (iii) Select the peptides with the lowest median of the five highest alignment scores (S_{A1} to S_{A5}). This approach may qualify as a robust alternative to the scores obtained by means of arithmetic means described above; it will not be affected by up to two errors in the labelling of non-allergen peptides.

Detection of proteins with allergenic potential using the ARP set

Using the established set of ARPs, the novel DASARP detector of potentially allergenic proteins was developed and evaluated. Briefly, each test protein was segmented, using a sliding window, into N_l peptides of the same length, l , as the ARPs. Subsequently, peptides from the test protein were compared with all ARPs using ungapped alignments. This results in N_a similarity (alignment) scores for each peptide and thus a N_l times N_a table of score values for the whole test protein.

This table of score values may be used for the detection in many different manners, but in the course of this work we have focused on the highest scores for each test peptide. Several different combinations of these scores were tested, but only the score combination yielding the best detection performance, is presented below. The detection (decision) statistic used is the sum ($S_{C1} + S_{C2}$) of the two largest elements, S_{C1} and S_{C2} , in the table (not belonging to the same test peptide), with a score above the detection statistic threshold indicating allergenicity. A flowchart describing the detection procedure is shown in Figure 2.

Evaluation of detector performance using 10-fold cross-validation

The performance of the novel method for the detection of allergenic potential, designated DASARP, was evaluated with a 10-fold cross-validation using one-tenth of all allergen sequences in each test set, whereas the remaining sequences were employed to obtain the ARPs. To avoid excessive repetition of the time-consuming comparison of allergen peptides to the large set of their non-allergen counterparts, this type of cross-validation was not applied to non-allergens. Hence, a set of non-allergen sequences was randomly extracted from the total repository and used for testing (naturally these test sequences were not employed in the procedure generating the ARP set).

The detector performance was evaluated using a novel form of receiver operating characteristic (ROC) curves. A conventional ROC curve depicts the probability of detection (true positive fraction, correctly classified allergens) on the ordinate versus the fraction of probability of false

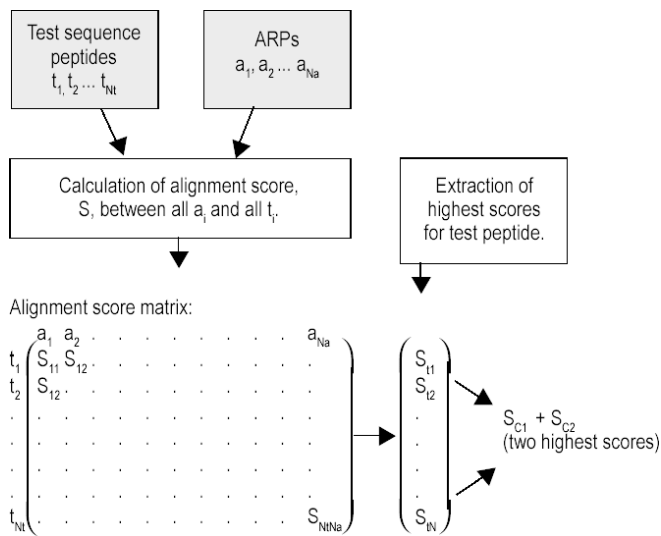


Fig. 2. A flowchart showing the principle behind detection of allergens based on similarity to the ARPs. First a test sequence is segmented into peptides of the same length as the ARPs and alignment scores between these peptides and all ARPs are calculated. Then the highest scores for each test peptide are extracted. The final detection statistic for the test protein is the sum of the two highest of these scores $S_{C1} + S_{C2}$, with higher value indicating a higher risk of allergenicity. Different detection statistic thresholds are reflected in the presented ROC curves.

alarm (false positive fraction, non-allergens erroneously classified as allergens) on the abscissa, as they vary with altered detection statistic thresholds, i.e. the trade-off between sensitivity and specificity for a two-class diagnostic test (Vining and Gladish, 1992). Owing to statistical variability, caused by the limited datasets available for the estimation of the true probability of detection (pDET) and the true probability of false alarm (pFA), an ROC curve unsupplied with this variation may become misleading. In a previous work, we introduced a new sort of performance curve, which includes variability (Soeria-Atmadja *et al.*, 2004). Here, it is refined into a more compact, informative form. The basic idea is simple: First, the confidence intervals for both quantities are determined by means of, e.g. K -fold cross-validation. Second, the point estimate of pDET is replaced by its lower confidence bound and the point estimate of pFA is replaced by its upper confidence bound, henceforth referred to as LBDET (lower bound pDET) and UBFA (upper bound pFA). This results in a more realistic ROC curve estimate that better reflects the true performance and therefore at least partly eliminates the risk of performance overstatement, compared to conventionally designed ROC curves.

Implementation

All computations were performed in the MATLAB™ programming environment (The MathWorks Inc.).

RESULTS

High-quality data repositories

A catalogue of 578 amino acid sequences from allergens was created as described in the System and methods section. A separate and comparatively large catalogue of amino acid sequences without documented association with allergy was built from commonly consumed food commodities. This rendered a non-allergen dataset with 22 176 protein sequences. Rice proteins were the predominant contributors to this dataset (18 812 sequences), since this is presently the most thoroughly sequenced species among the commodities used.

Detector performance

Several methods to accomplish a specific selection of the ARPs were tested, as detailed in the System and methods section. The overall performance was measured by testing with 57/58 allergens and 700 non-allergens, for each 10 DASARP cross-validation cycles. The ARP set in each cycle was derived with ASARP, using the 520/521 allergens and 22 176 – 700 = 21 476 non-allergens. The respective detection rates (LBDET) of these methods at 10, 15 and 20% false classification of non-allergens (UBFA), using a peptide length of 24 amino acids ($l = 24$) and the Blosum80 substitution matrix, are summarized in Table 1. Several values of the threshold T_A were evaluated but only those returning the best results (highest LBDET at the fixed UBFA level) are shown. The best performance result was obtained with the mean value of the 10 highest alignment scores as the decision statistic for ARP selection. Two substitution matrices, Blosum80 and Pam30, both allegedly suited for alignment of short sequences (Altschul, 1991), were compared using a single matrix in both key steps, selection of ARPs and classification of test sequence peptides. Blosum80 consistently outperformed Pam30 for all methods described above. With longer peptides, however, detector performance with Pam30 proved to be nearly as good as that of Blosum80. Owing to an overall preference for Blosum80, all results presented below are based on this substitution matrix.

Figure 3 shows the new performance curve for DASARP, employing $S_{C1} + S_{C2}$ as the detection statistic and selecting ARPs with the mean of the 10 highest alignment scores. To provide for practical applications, we defined the best performing algorithm as that yielding the highest LBDET when the UBFA was set to 10%. This approach resulted in a LBDET of (at least) 81%, corresponding to a threshold score for $S_{C1} + S_{C2}$ at 5.51.

Since the number of critical features (causing the allergic reactions) of an average allergen is not well known, there are intrinsic difficulties as to the determination of the optimal number of ARPs in advance. Accordingly, several values for n_a (number of ARPs per allergen), ranging from 1 to 20, were examined. Judged from such tests, five ARPs proved most successful for detection performance (data not shown). All

Table 1. Minimum rate of allergen detection (LBDET) given a desired maximum rate of false alarms (UBFA) set at three different UBFA levels: 10, 15 and 20%

Similarity score	UBFA (%) 95% conf.	LBDET (%) 95% conf.	Mean pFA (%)	Mean pDET (%)
n_h	9.5	80.9	8.5	89.3
	14.2	81.2	12.7	89.6
	18.7	80.6	17.3	90.7
Mean ($S_{A1}, S_{A2}, \dots, S_{A10}$)	9.9	81.2	9.2	89.1
	14.3	82.1	13.3	89.4
	19.1	83.7	18.0	90.5
Mean (S_{A1}, S_{A2}, S_{A3})	9.6	79.1	8.8	88.8
	14.9	81.0	13.9	89.9
	18.6	82.7	17.1	90.8
Median ($S_{A1}, S_{A2}, \dots, S_{A3}$)	9.6	80.2	9.0	89.4
	14.1	81.6	13.1	90.7
	17.8	82.7	16.8	91.3

These values were evaluated for four different methods for selection of ARPs, all with peptides of length 24 amino acids. Tests were performed with a 10-fold cross-validation using $S_{C1} + S_{C2}$ as the detection statistic. The average (mean pDET and mean pFA, equivalent to conventional ROC-curves) are also included in the table.

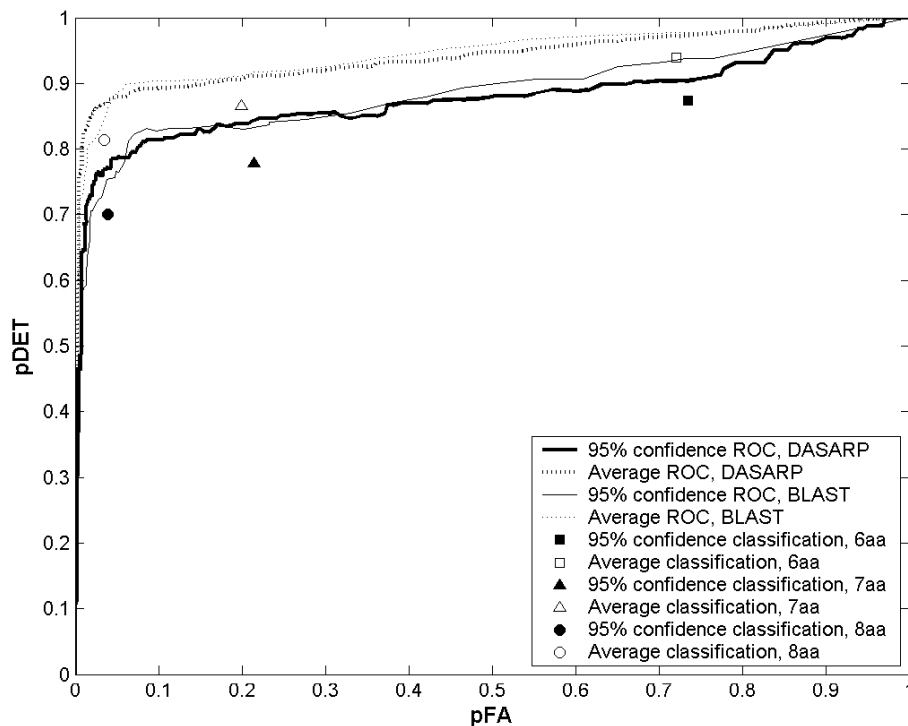


Fig. 3. Conservative ROC curve for DASARP using ARPs of peptide length 24, selected with lowest mean ($S_{A1}, S_{A2}, \dots, S_{A10}$) compared with detector using BLAST local alignment over an 80 amino acid window. For both methods, detector performance with 95% confidence as well as average performance is presented. The conservative ROC curve for the BLAST procedure was generated by incrementing the sequence identity threshold from 0 to 100%. Detection using 6, 7 or 8 consecutive identical amino acids as determinant of allergenicity are inserted as three separate points together with the average performance for these three tests. Exact values for these last tests can be found in Table 3.

ASARP parameter sets were evaluated in tests, using peptide lengths ranging from 6 to 36 amino acids. DASARP performance values based on some of the evaluated peptide lengths are listed in Table 2. The best detector results (as defined above) were obtained with a peptide length of 24 amino acids.

Computational issues

On a standard 1 GHz processor, one round of ASARP with the calculation of alignment score between all peptides lasted 2–3 days and subsequent detection with 10-fold cross-validation, one more day. This was done for several different peptide

Table 2. Minimum rate of allergen detection (LBDET) given a desired maximum rate of false alarms (UBFA) as well as average values for DASARP with three different peptide lengths

Peptide length	UBFA (%) 95% conf.	LBDET (%) 95% conf.	Mean pFA (%)	Mean pDET (%)
18	8.7	76.8	7.6	87.5
	13.6	78.3	12.2	88.9
	18.7	79.7	17.2	90.0
24	9.9	81.2	9.2	89.1
	14.3	82.1	13.3	89.4
	19.1	83.7	18.0	90.5
30	9.2	79.6	8.4	88.6
	13.8	81.7	12.9	90.0
	17.8	82.7	16.6	90.8

The detection procedure was performed using ARPs selected with lowest mean (S_{A1} , S_{A2} , ..., S_{A10}) and $S_{C1} + S_{C2}$ as detection statistic.

lengths as well as with different parameter settings. One test set containing 700 non-allergen sequences was employed.

A set of allergen-specific peptides

The ASARP procedures rendered a final set of five peptides from each of the 578 allergens (in total 2890 ARPs), with the purpose of detecting amino acid sequences of potential protein allergens. To reveal whether the ARPs are widely scattered throughout the sequences from which they are derived, or largely restricted to narrow regions only, the degree of ARP overlap was studied. Figure 4 shows the distribution across all 578 allergens. In a few cases, especially for the longer proteins, the ARPs are not overlapping at all. For most allergens, however, partial (but not extensive) overlaps occur. Hence, certain parts of the entire set of allergen amino acid sequences have preferential coverage and are thus likely to be especially important.

To investigate whether a functional connection to allergenic properties contributed to this finding we compared the ARPs with known IgE and T-cell epitopes. In Figure 5 two examples of allergen sequences (Par j 2 and Asp f 2), including ARPs and known epitopes, are shown. The IgE-epitopes were mined (April 2004) from SDAP (Ivanciuc *et al.*, 2003), whereas the T-cell epitopes for Asp f 2 were retrieved from Svirshchevskaya *et al.* (2002). Even though the ARPs cover epitope motifs in many cases, superimposed positions were not obvious. A roughly similar pattern was evident in an extended set of ARP/epitope location maps of arbitrarily chosen amino acid sequences of allergen proteins (data not shown).

Prediction of allergenicity according to FAO/WHO recommendations

As mentioned above, a match of six consecutive amino acids to known allergens, is suggested to indicate potential allergenicity of the query sequence, according to the first criterion

of FAO/WHO recommendations (FAO/WHO, 2001). Other reports suggest stretches of 7 or 8 amino acids as the critical limit (Goodman *et al.*, 2002; Metcalfe *et al.*, 1996). Hence, we compared the detection performance based on identity over 6, 7 or 8 amino acids with our new DASARP method. The evaluation was conducted using the same datasets mentioned above, with 10-fold cross-validation for allergens, and the test set for non-allergens with 700 sequences. The results from these tests are shown in Figure 3 and Table 3.

The second criterion in the bioinformatics allergenicity-testing scheme recommended by FAO/WHO is to segment the query sequence using a sliding window of 80 amino acids and subsequently align these peptides against an allergen database, using the local alignment algorithm FASTA (Pearson, 2000). If any of the 80 amino acids long peptides has more than 35% sequence identity with any of the allergens in the database the query sequence shall be assigned as allergenic (FAO/WHO, 2001). This evaluation is, however, not well defined as to parameter-settings such as substitution matrices, gap penalties, etc. Owing to more convenient computational access to the local alignment algorithm BLAST (Altschul *et al.*, 1990), the above-mentioned test procedure was evaluated using BLAST (standard settings with the Blosum62 substitution matrix) instead of FASTA. This method was cross-validated using the above-mentioned datasets. All amino acid sequences of the test sets were segmented into 80 residues long fragments, except for those of shorter length. From BLAST alignments percentage identity scores were calculated for matches along the query sequence length. The results were compared with those of DASARP and the identical peptide match method, using conservative ROC curves generated by incrementing the sequence identity threshold from 0 to 100% (Fig. 3). The BLAST alignment performance at 35% sequence identity threshold is also summarized in Table 3.

DISCUSSION

Using properties of allergenic proteins to design a detector of allergenic potential

Several characteristics are commonly associated with proteinaceous food allergens. Examples thereof are heat stability, resistance to protease digestion in acidic environment, relatively large molecular weight and carbohydrate modification (Bredehorst and David, 2001; Huby *et al.*, 2000). Proteins not considered to be allergenic, however, can possess any of the above-listed features. Data from structural biochemistry and computer searches reveal an extensive diversity among hypersensitivity type I-associated proteins, as well as within the smaller group food allergens (Aalberse, 2000; Bredehorst and David, 2001). Nonetheless, the main repository of allergenic proteins falls within a fairly limited range of protein families, but most members of such families are still not allergenic although sharing an overall structural outline with one or several allergens (Aalberse *et al.*, 2001; Mills *et al.*, 2003).

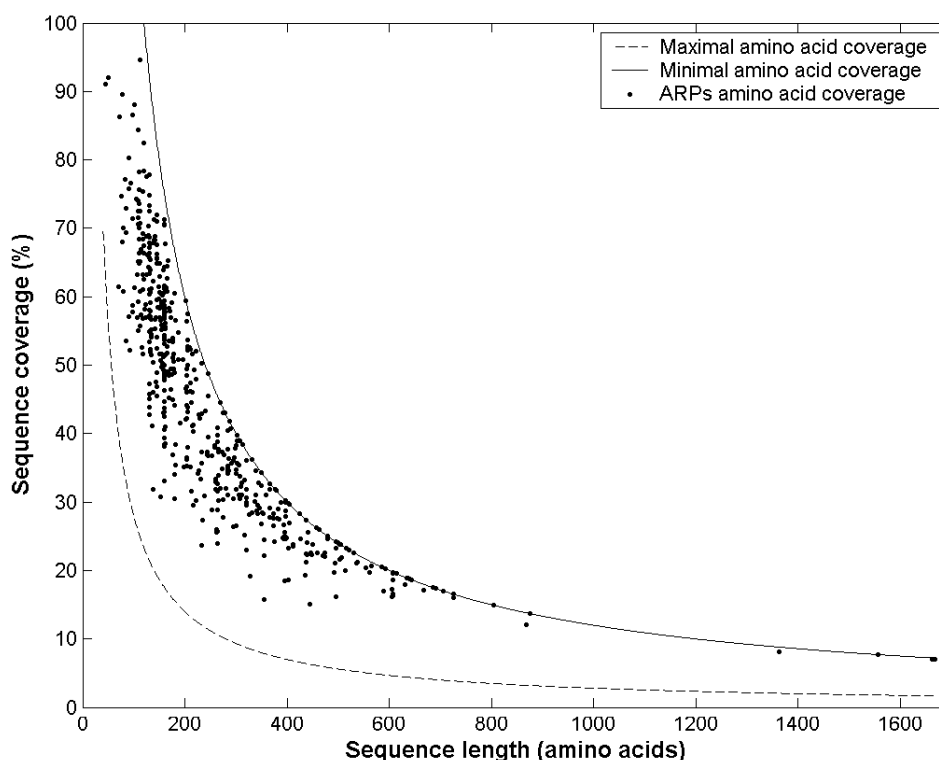


Fig. 4. Amino acid coverage of ARPs. The points show how much (amino acid percentage) each allergen sequence is covered by its corresponding ARPs. The figure also includes a curve showing the maximal amino acid coverage (lowest possible overlap between ARPs) as well as a curve showing the minimal amino acid coverage (full overlap between all five ARPs) at a given sequence length.

This highlights a key problem associated with the design of methods with good detection in conjunction with few false alarms: conserved sequences in proteins may not be related to their allergenic potential. Algorithms solely focusing on conserved motifs in allergens give good detection in many instances, but since they have an inherent tendency to assign non-allergens from structure-similar families as allergens, these methods are not easily improved towards fewer false positives. To tackle this problem, we have developed a novel concept, built on dissimilarities between well-documented allergenic proteins and those with a high probability of being devoid of this feature. This computational approach, named DASARP, is founded on a preferential elimination of motifs in allergens that commonly occur in its non-allergen counterparts.

In our previous learning systems approach to the detection of allergen potential, each test protein sequence was aligned to a set of pre-selected prototype allergen sequences, and features such as the alignment length and the alignment score were extracted. These features were then merged into a characteristic pattern (analogous to a fingerprint), which was allowed to feed a supervised classification algorithm (Soeria-Atmadja *et al.*, 2004; Zorzet *et al.*, 2002). In the new approach, DASARP, a test protein is decomposed into peptides that are individually compared with the ARPs collected. Then the

score values of the two best matches are directly merged into a single feature, which is used for classification. In other words, no characteristic pattern that consists of several features is created, and there is no supervised classification algorithms involved. The results reported here show that the new features extracted using the ARPs are powerful enough to allow high-standard detection of allergenic potential simply by using the sum of two particular features selected based on a biological rationale. Plausibly, the use of more complex combinations with more than two features would yield even better detector performance. Therefore, in our future work, we plan to employ supervised classification algorithms in the intricate search for such combinations.

DASARP shows good performance in comparison with previous methods

Taking statistical variations into account, the best detector developed here showed an LBDET of 81% when allowing for a probability of false alarm up to 10% (UBFA). In one of our previous reports, an alignment-based learning system for the detection of potential allergens is described (Soeria-Atmadja *et al.*, 2004), allowing an LBDET of 70% at 11% UBFA. Since a smaller set of allergens was used in that study the performance of the method is, however, not directly comparable

NL21_PARJU (Par j 2)
MRTVSMAALVVIAAALAWTSSAEPAPAPAPGEEACGKVVQDIMPCLHFKGEEKEPSKECCS
MRTVSMAALVVIAAALAWTSSAEPAPAPAPGEEACGKVVQDIMPCLHFKGEEKEPSKECCS

GTKKLSEEVKTTEQKREACKCIVRATKGISGIKNELVAEVPKKCDIKTTLPITADFDCSKI
GTKKLSEEVKTTEQKREACKCIVRATKGISGIKNELVAEVPKKCDIKTTLPITADFDCSKI

QSTIFRGYY
QSTIFRGYY

ALL2_ASPFU (Asp f 2)
MAALLRLAVLLPLAAPLVATLPTSPVPIAARATPHEPVFFSWDAGAVTSFPIHSSCNATQRR
MAALLRLAVLLPLAAPLVATLPTSPVPIAARATPHEPVFFSWDAGAVTSFPIHSSCNATQRR
MAALLRLAVLLPLAAPLVATLPTSPVPIAARATPHEPVFFSWDAGAVTSFPIHSSCNATQRR

QIEAGLNEAVELARHAKAHILRWGNESEIYRKYFGNRPTMEAVGAYDIVNGDKANVLFRC
QIEAGLNEAVELARHAKAHILRWGNESEIYRKYFGNRPTMEAVGAYDIVNGDKANVLFRC
QIEAGLNEAVELARHAKAHILRWGNESEIYRKYFGNRPTMEAVGAYDIVNGDKANVLFRC

NPDGNCALEGWGGHWRGANATSETVICDRSYTTRRWLVSMCSQGYTVAGSETNTFWASDLMH
NPDGNCALEGWGGHWRGANATSETVICDRSYTTRRWLVSMCSQGYTVAGSETNTFWASDLMH
NPDGNCALEGWGGHWRGANATSETVICDRSYTTRRWLVSMCSQGYTVAGSETNTFWASDLMH

RLYHVPVAVGQGWVDHFADGYDEVIALAKSNGTESTHDSEAFYFALEAYAFDIAAPGVGCAG
RLYHVPVAVGQGWVDHFADGYDEVIALAKSNGTESTHDSEAFYFALEAYAFDIAAPGVGCAG
RLYHVPVAVGQGWVDHFADGYDEVIALAKSNGTESTHDSEAFYFALEAYAFDIAAPGVGCAG

ESHGPDQGHDTGSASAPASTSTSSSSSGSGGATTTPTDSPSATIDVPSNCHTHEGGQLHCT
ESHGPDQGHDTGSASAPASTSTSSSSSGSGGATTTPTDSPSATIDVPSNCHTHEGGQLHCT
ESHGPDQGHDTGSASAPASTSTSSSSSGSGGATTTPTDSPSATIDVPSNCHTHEGGQLHCT

Fig. 5. Amino acid sequences of the allergens Par j 2 and Asp f 2. The underscored regions of the first line of each text block show where the ARPs are situated whereas in the second line they represent documented IgE-epitopes. In Asp f2 there is a third line where the underscored segments correspond to documented T-cell epitopes.

Table 3. Minimum rate of allergen detection (LBDET) and maximum rate of false alarms (UBFA) for classifications using 6, 7 and 8 consecutive identical amino acids

Peptide length	UBFA (%) 95% conf.	LBDET (%) 95% conf.	Mean pFA (%)	Mean pDET (%)
6	73.4	87.2	72.0	94.0
7	21.4	77.7	19.9	86.5
8	3.9	69.8	3.5	81.3
BLAST	7.3	82.4	5.8	89.3

Detection performance based on 35% homology (as revealed by BLAST) over 80 amino acids is also listed. Average values are also included.

with that presented in this paper. Because of the inherent differences between alignment of entire amino acid sequences and DASARP, it is likely that each of them possesses unique advantages. Hence, we are presently considering the possibilities of merging both strategies into a single detector of allergenic potential.

The aforementioned bioinformatic assessment of allergenic potential recommended by the FAO/WHO, via a search for identical matches of six consecutive amino acids (FAO/WHO, 2001), provides for a good detection level though at the

expense of a very high false alarms rate. Our method outperforms detection founded on this principle, as well as ones based on the recognition of 7 or 8 amino acid matches. As mentioned above, the FAO/WHO recommendations include an alternative procedure based on alignment over an 80 amino acid window using sequence identity >35% to indicate allergic cross-reactivity (FAO/WHO, 2001). This procedure was carried out and compared to DASARP. ROC curves in Figure 3 reveal comparable results on the performance for DASARP and BLAST alignment.

DASARP, using the detection statistic threshold we found most appropriate (5.51), detected—in most cases—the same allergens as those targeted by BLAST alignment with a cut-off limit of 35%. Some allergens were, however, recognized by one method but overlooked by the other (Table 4). We believe, though, a more flexible ARP selection procedure is the key to DASARP enhancement, presumably leading to greater differences in targeted amino acid sequences, relative to the 35% sequence identity (over an 80 amino acid window) approach.

Influence of non-allergen data

A wide-ranged set of non-allergenic food proteins was used to develop the DASARP detector. This dataset is large and covers several protein families. Still, it does not contain all

Table 4. Allergens picked up by DASARP but that were overlooked with BLAST alignments over an 80 amino acids window and vice versa

Allergens identified by DASARP but not by BLAST	DASARP score value. (detection statistic threshold is 5.51)	Allergens identified by BLAST but not by DASARP	Sequence identity (%) (detection statistic threshold is 35%)
ALL2_CANFA	5.91	E13B_HEVBR	47.5
Q8W3Y9	7.25	LCA_BOVIN	45
MPT5_AMBTR	6.87	GTM1_DERPT	39
ALL2_ASPFU	5.83	Q9AVK8	36
AAO24900	6	ALL7_LEPDS	36
Q9U3U5	7.16	OVAL_CHICK	37.5
AMYG_ASPNG	5.66		

The detection statistic threshold (score value or sequence identity) for each allergen is also listed.

possible structures of non-allergenic proteins. Hence, some allergens may have few, if any, structural counterparts in the set. The ASARP method is aimed at finding sub-sequences that distinguishes the allergens from its non-allergenic counterparts, but due to insufficiencies in the non-allergen dataset, the selection of ARPs might have, at least in some cases, been biased towards peptides without counterparts in our set of non-allergen sequences. It may therefore prove rewarding to perform the ASARP on allergens with a more carefully selected set of non-allergens, designed to compensate for possible imbalances. There are, however, clear difficulties entailed in designing a large non-allergen set without contamination of allergens, since documentation of non-allergenicity is available only for a few selected proteins.

Does a peptide length of 24 amino acids indicate the size of an allergenic motif?

Many different peptide lengths, ranging from 6 to 36 amino acids, were tested for DASARP detection performance. The highest detection rates (at a fixed level of false alarms) were consistently obtained with a peptide length of 24 amino acids. Some ARPs were found to cover reported motifs, such as T-cell and IgE epitopes, although no clear correspondence between ARPs and epitopes was seen. Putatively, this is due to variable motif length and different motif numbers within and across the allergens, respectively.

An ASARP algorithm that affords flexible selection of peptides as regards length and number, would theoretically be better to target functional motifs selectively, relative to settings used in this paper. Such an approach would, however, drastically increase the computational complexity. Since our primary incentive is to provide a detection method (of allergenic potential) for safety assessment purposes and to demonstrate the potential of the ASARP process as such, detection founded on variable ARP lengths is a topic beyond the scope of the present study. A separate but related issue involves gap-opening penalty setting in the peptide comparison procedure. In its present design, our algorithm does not allow for gaps because discontinuous sequence alignments would extend the computational processing time considerably. Nonetheless, alignment gaps

would open for selection of larger motifs, and may lead to more accurate selection of sequences.

Statistical variability in our results

Although the results reported here take statistical variability into account, it is important to stress the inherent difficulties associated with estimation of the probabilities of detection and false alarm that ultimately appear as uncertainties, e.g. in the seemingly conservative ROC curves presented. In particular, the use of 10-fold cross-validation for estimation of the average and variance of the probability of detection (pDET) have several implications. First of all, since only 10 point estimates are computed (one for each fold), the variance estimate is uncertain. From this can be inferred that the confidence intervals should be computed assuming the Student's-*t* rather than the normal (Gaussian) probability density function. Second, the accuracy of the point estimates are somewhat compromised by the low numbers of allergen test examples (57 or 58) used in each iteration. For example, in average 6 errors occurred among 57 allergens in our tests. This implies a point estimate of ~11% for the error rate (pDET), whereas the corresponding 95% Bayesian credible interval of the error rate is [4.3%, 20.2%]. Similarly, with 63 errors in 700 non-allergen examples obtained, the credible interval for the error rate was [7.0%, 11.3%]. Such relatively large uncertainties associated with detector performance validation were recently discussed also in the context of classification and detection of cancers (Simon *et al.*, 2003). Third, *K*-fold cross-validation does not result in a completely natural variability in the design sets. This is most easily seen in the special case of leave-one-out cross-validation where all detectors are designed with almost identical design sets. In conclusion, although we have taken measures to reflect the statistical variability in the results presented here, there are additional sources of variability, which should be kept in mind when interpreting and comparing present and past results.

CONCLUSION

DASARP is unique in its kind since it is founded on preferential reduction of motifs in amino acid sequences of protein

allergens, which occur more commonly in proteins devoid of either (type-I hypersensitivity-) sensitizing or triggering features. No biocomputational method for the detection of allergenic potential reported elsewhere exploits this opportunity. Instead, current methods are focused on inter-allergen similarities. By depleting protein allergen amino acid sequences of motifs common among all proteins, sets of allergen-representative peptides, ARPs, were selected. These peptides were then used to detect potential allergenicity using a simple decision procedure. The novel detector outperformed and compared well with the identical peptide match method and the alignment-based detection procedure, respectively, both outlined and recommended by FAO/WHO (2001). DASARP, however, shows certain unique amino acid sequence targeting features and has inherent potential for refinement.

ACKNOWLEDGEMENTS

We thank Ms. Maj Olausson for expert assistance on computer graphics. Financial support for this work was provided by the Swedish Agency for Innovation Systems (VINNOVA), Carl Tryggers Stiftelse (Stockholm), Göran Gustafssons Stiftelse (Stockholm), the Faculty of Science and Technology (Uppsala University), the cancer and allergy fund and the Swedish National Board for Laboratory Animals (CFN).

REFERENCES

- Aalberse, R.C. (2000) Structural biology of allergens. *J. Allergy Clin. Immunol.*, **106**, 228–238.
- Aalberse, R.C., Akkerdaas, J. and van Ree, R. (2001) Cross-reactivity of IgE antibodies to allergens. *Allergy*, **56**, 478–490.
- Altschul, S.F. (1991) Amino acid substitution matrices from an information theoretic perspective. *J. Mol. Biol.*, **219**, 555–565.
- Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
- Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M.C., Estreicher, A., Gasteiger, E., Martin, M.J., Michoud, K., O'Donovan, C., Phan, I., Pilbout, S. and Schneider, M. (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.*, **31**, 365–370.
- Brederhorst, R. and David, K. (2001) What establishes a protein as an allergen? *J. Chromatogr. B Biomed. Sci. Appl.*, **756**, 33–40.
- Brusic, V., Millot, M., Petrovsky, N., Gendel, S.M., Gigonzac, O. and Stelman, S.J. (2003) Allergen databases. *Allergy*, **58**, 1093–1100.
- Chuang, V.T., Kragh-Hansen, U. and Otagiri, M. (2002) Pharmaceutical strategies utilizing recombinant human serum albumin. *Pharm. Res.*, **19**, 569–577.
- FAO/WHO (2001) Evaluation of allergenicity of genetically modified foods. Joint FAO/WHO Expert Consultation on Allergenicity of Foods Derived from Biotechnology.
- FAO/WHO (2003) Codex principles and guidelines on foods derived from biotechnology, Codex Alimentarius. Joint FAO/WHO Food Standards Programme.
- Gasteiger, E., Gattiker, A., Hoogland, C., Ivanyi, I., Appel, R.D. and Bairoch, A. (2003) ExPASy: the proteomics server for in-depth protein knowledge and analysis. *Nucleic Acids Res.*, **31**, 3784–3788.
- Gendel, S.M. (1998a) Sequence databases for assessing the potential allergenicity of proteins used in transgenic foods. *Adv. Food Nutr. Res.*, **42**, 63–92.
- Gendel, S.M. (1998b) 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.
- Gendel, S.M. (2002) Sequence analysis for assessing potential allergenicity. *Ann. N. Y. Acad. Sci.*, **964**, 87–98.
- Goodman, R.E., Silvanovich, A., Hileman, R.E., Bannon, G.A., Rice, E.A. and Astwood, J.D. (2002) Bioinformatic methods for identifying known or potential allergens in the safety assessment of genetically modified crops. *Comments Toxicol.*, **8**, 251–269.
- Gualandi-Signorini, A.M. and Giorgi, G. (2001) Insulin formulations—a review. *Eur. Rev. Med. Pharmacol. Sci.*, **5**, 73–83.
- Heppenheimer, T.A. (2003) The growth of genetically modified foods. *Am. Herit. Invent. Technol.*, **19**, 16–25.
- 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.
- Huby, R.D., Dearman, R.J. and Kimber, I. (2000) Why are some proteins allergens? *Toxicol. Sci.*, **55**, 235–246.
- Ivanciuc, O., Schein, C.H. and Braun, W. (2002) Data mining of sequences and 3D structures of allergenic proteins. *Bioinformatics*, **18**, 1358–1364.
- Ivanciuc, O., Schein, C.H. and Braun, W. (2003) SDAP: database and computational tools for allergenic proteins. *Nucleic Acids Res.*, **31**, 359–362.
- Jansen, J.J., Kardinaal, A.F., Huijbers, G., Vlieg-Boerstra, B.J., Martens, B.P. and Ockhuizen, T. (1994) Prevalence of food allergy and intolerance in the adult Dutch population. *J. Allergy Clin. Immunol.*, **93**, 446–456.
- Kanny, G., Moneret-Vautrin, D.A., Flabbee, J., Beaudouin, E., Morisset, M. and Thevenin, F. (2001) Population study of food allergy in France. *J. Allergy Clin. Immunol.*, **108**, 133–140.
- King, T.P., Hoffman, D., Lowenstein, H., Marsh, D.G., Platts-Mills, T.A. and Thomas, W. (1995) Allergen nomenclature. *Allergy*, **50**, 765–774.
- 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.
- Kuehn, B.M. (2003) Bioengineered pigs go to market. *J. Am. Vet. Med. Assoc.*, **222**, 926.
- Lee, Y.H. and Sinko, P.J. (2000) Oral delivery of salmon calcitonin. *Adv. Drug Deliv. Rev.*, **42**, 225–238.
- 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.
- Mills, E.N.C., Madsen, C., Shewry, P.R. and Wichers, H.J. (2003) Food allergens of plant origin—their molecular and evolutionary relationships. *Trends Food Sci. Technol.*, **14**, 145–156.
- Pearson, W.R. (2000) Flexible sequence similarity searching with the FASTA3 program package. *Methods Mol. Biol.*, **132**, 185–219.

- SCP (1998) The occurrence of severe food allergies in the EU. Report of experts participating in task 7.2, Scientific Cooperation Programme in the EU.
- Simon,R., Radmacher,M.D., Dobbin,K. and McShane,L.M. (2003) Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification. *J. Natl Cancer Inst.*, **95**, 14–18.
- 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.
- Soltero,R. and Ekwuribe,N. (2002) The oral delivery of protein and peptide drugs. *Innovat. Pharmaceut. Technol.*, **1**, 106–110.
- Stadler,M.B. and Stadler,B.M. (2003) Allergenicity prediction by protein sequence. *FASEB J.*, **17**, 1141–1143.
- Svirshchetskaya,E.V., Alekseeva,L., Marchenko,A., Viskova,N., Andronova,T.M., Benevolenskii,S.V. and Kurup,V.P. (2002) Immune response modulation by recombinant peptides expressed in virus-like particles. *Clin. Exp. Immunol.*, **127**, 199–205.
- Vining,D.J. and Gladish,G.W. (1992) Receiver operating characteristic curves: a basic understanding. *Radiographics*, **12**, 1147–1154.
- Zorzet,A., Gustafsson,M. and Hammerling,U. (2002) Prediction of food protein allergenicity: a bioinformatic learning systems approach. *In Silico Biol.*, **2**, 525–534.