

Statistical Evaluation of Local Alignment Features Predicting Allergenicity Using Supervised Classification Algorithms

D. Soeria-Atmadja^a A. Zorzet^a M.G. Gustafsson^b U. Hammerling^a

^aDivision of Toxicology, National Food Administration, and ^bSignal and Systems Group, Uppsala University, Uppsala, Sweden

Key Words

Allergy · Amino acid sequence · Computational toxicology · Risk assessment

Abstract

Background: Recently, two promising alignment-based features predicting food allergenicity using the k nearest neighbor (kNN) classifier were reported. These features are the alignment score and alignment length of the best local alignment obtained in a database of known allergen sequences. **Methods:** In the work reported here a much more comprehensive statistical evaluation of the potential of these features was performed, this time for the prediction of allergenicity in general. The evaluation consisted of the following four key components. (1) A new high quality database consisting of 318 carefully selected, non-redundant allergens and 1,007 sequences carefully selected to be non-allergens. (2) Three different supervised algorithms: the kNN classifier, the Bayesian linear Gaussian classifier, and the Bayesian quadratic Gaussian classifier. (3) A large set of local alignment procedures defined using the FASTA3 alignment program by means of a wide range of different parameter settings. (4) Novel performance curves, alternative to conventional receiver-operating characteristic curves, to display not only average behaviors but also statistical variations due to small data sets. **Results:** The linear Gaussian classifier proved most useful among the tested super-

vised machine learning algorithms, closely followed by the quadratic Gaussian equivalent and kNN. The overall best classification results were obtained with a novel feature vector consisting of the combined alignment scores derived from local alignment procedures using different substitution matrices. **Conclusions:** The models reported here should be useful as a part of an integrated assessment scheme for potential protein allergenicity and for future comparisons with alternative bioinformatic approaches.

Copyright © 2004 S. Karger AG, Basel

Introduction

Atopic allergy and other hypersensitivity reactions affect up to 25% of the population in affluent industrial nations. The prevalence of food allergenicity among adults within the EU ranges from 0.8 to 2.4% and is even higher in the pediatric population [1–3]. More than 90% of allergic food hypersensitivity are accounted for by eight foods, of which peanuts, soybeans, tree nuts and crustaceans comprise the most commonly involved sources [4]. Agricultural biotechnology enables the introduction of specific genes into food crops. A common utilization of this technique involves the expression of xenogenic proteins, which for example confer resistance to herbicide, various plant pests or plant viruses. Consumer safety issues in the context of genetically modified (GM) foods

KARGER

Fax +41 61 306 12 34
E-Mail karger@karger.ch
www.karger.com

© 2004 S. Karger AG, Basel
1018–2438/04/1332–0101\$21.00/0

Accessible online at:
www.karger.com/iaa

Correspondence to: M.G. Gustafsson
Signal and Systems Group, Uppsala University
PO Box 528
SE-751 20 Uppsala (Sweden)
Tel. +46 18 471 32 29, E-Mail mg@signal.uu.se

have been highlighted by several organizations. These include the European Commission, the Joint International Life Science Institute – International Food Biotechnology Council (ILSI/IFBC) and the Joint Food and Agriculture Organization and World Health Organization Expert Consultation on Foods Derived from Biotechnology [5–7]. Particularly the risk of exposure to potential allergens has attracted considerable attention. The inadvertent introduction of allergenic properties by transgene proteins can emanate either from sources with a known offending capacity or from organisms devoid of a history of safe consumption. Moreover, transgenes can in principle also alter the host's expression of endogenous allergens [8].

A decision tree for the safety assessment of potential allergenicity of GM foods was presented by the ILSI/IFBC in 1996 [5]. This scheme involves a range of test procedures, e.g. an immunosorbent assay, resistance to digestion, skin prick tests and food challenges. The path taken through the program depends on test outcomes. Notably, a comparison of the amino acid sequence of the novel protein with those of known allergens is suggested as an introductory step for transgenes devoid of a history of allergy. A sequence match of eight identical consecutive amino acids with a known allergen sequence is proposed in the above-mentioned scheme to raise the alarm for potential offending capacity [5]. Subsequently, the joint FAO/WHO Expert Consultation on Allergenicity of Foods Derived from Biotechnology has advocated a revised decision tree for the assessment of potential GM food allergenicity. In this scheme, interrogation using amino acid sequence alignment, based on the FASTA program, is mandatory to all test proteins, regardless of the source [6]. A changed threshold for alarm to six consecutive (identical) amino acids was suggested in the aforementioned report. Moreover, a second criterion, specifying an identity of more than 35% over an 80-amino acid window, was proposed. Either of the two above-specified criteria is suggested to indicate potential sensitizing property or cross-reactivity [6].

There are several additional reports on local alignment setting preferences for safety assessment of novel proteins in the food chain, especially as expressed by GM crops. Gendel [9, 10] has suggested a bifurcated strategy, involving an initial search for matches of identical amino acids, followed by inspection using either a biochemical or an evolutionary scoring matrix. A frequent occurrence of five or six consecutive identical amino acids by chance has been shown [9, 11, 12]. Furthermore, few cross-reactive allergens have a sequence similarity below 50%

across the entire length of the protein sequences and 70% or more is a typical range [13]. Nonetheless, percentage overall similarity may lead astray: shrimp tropomyosin, a well known allergen, shows about 60% sequence similarity to that of chicken, but vertebrate tropomyosins are not known to trigger hypersensitivity reactions in humans [14].

Other proposed bioinformatic approaches to classify amino acid sequences in the context of allergenicity involve comparison of a test sequence with available data on IgE epitopes. This course has been suggested either as a stand-alone method or as an extension to the above-mentioned test/allergen alignment procedure, using a window setting of six amino acids [12, 15]. In one of these settings, a search scheme for either sequence identity or matching properties based on peptide physicochemical features is included [15]. Recently, a classification method on allergenicity, founded on the MEME motif discovery tool [16], was reported [17]. The performance on classification of allergenic proteins was reported as being comparable with that of the FAO/WHO procedure, but much higher with respect to non-allergens [17].

In this paper, we present a major extension of an earlier work on a supervised learning system which was based on the *k* nearest neighbor (kNN) classifier algorithm combined with local alignments to recognize sensitizing food allergens only [18]. Here, three different classifier algorithms are used in combination with a much larger set of different local alignments procedures obtained with the FASTA3 program. In each case, a pair of extracted FASTA3 output features was feeding three different conventional supervised classifier algorithms; the kNN classifier, the Bayesian linear Gaussian classifier (LG) and the Bayesian quadratic Gaussian classifier (QG) [19, 20]. Furthermore, a much more careful statistical evaluation of the different classifier systems, designed by combining a classifier algorithm with a particular local alignment parameter setting, was performed. Five separate scoring matrices and two distinct gap opening and extension penalty settings were employed. The statistical evaluation involved 200 redesigns of the classifier systems using 70 allergen and 272 non-allergen randomly selected test examples for performance evaluation each time. The best results were compared by means of a new kind of performance curve, for the first time introduced here, which should be regarded as an alternative to the conventional receiver-operating characteristic (ROC) curve [21]. The new curves not only include the average performance but also the statistical variation caused by the relatively small data sets used.

In addition to a markedly refined statistical evaluation and a much larger set of classifier systems considered here, relative to our earlier work, a new and expanded in-house database was employed. It consists of 318 allergen and 1,007 non-allergen amino acid sequences, each subcatalogue representing a carefully selected and non-redundant excerpt from several publicly available repositories. The results indicate that detection of allergen sequences based on local alignment scores and alignment length only may not be the best possible approach but nonetheless useful as a risk assessment tool. Moreover, the work is an important reference to which any alternative procedure proposed for the detection of allergen sequences should be compared.

Materials and Methods

Database Mining and Establishment of In-House Repositories

Amino acid sequences of allergen proteins were mined from five different publicly available databases: Allergen nomenclature [22]; Farrp [23]; the Allergen Database [24]; the Allergen Sequence Database [25], and the Protall Database [26]. Prior to filing in our in-house repository, each record was inspected for presence in at least two of the aforementioned catalogues. Partial sequences shorter than 200 amino acids were omitted. In total 318 records were allocated to the total allergen set, which is essentially deliberated from isoallergens. A file of 3,000 amino acid sequences unlikely to be associated with allergy was created by searching the SWall database (SwissProt and TrEMBL) [27], using SRS (Sequence Retrieval System) at the ExPASy (Expert Protein Analysis System) proteomics server of the Swiss Institute of Bioinformatics (SIB) [28]. Entries holding the text strings 'allergen', 'atopy' or derivatives thereof in the annotations were excluded. For practical reasons the primary data set of 3,000 sequences was reduced to 1,007. To minimize the bias of the distribution with regard to protein structure, this reduction procedure was accomplished with the aid of multiple alignment using ClustalX [29]. This non-allergen repository has the following species representation: Apple: 100, carrot: 75, peach, cherry or apricot: 185, spinach: 105, tomato: 380, salmon: 105, cow's milk: 46 and hen's egg: 11. Listings of protein accession assignments of the above-mentioned data records, both allergenic and non-allergenic, are available at the following site: <http://www.slv.se/allergyalign/tables>.

Computer Programming

Except for FASTA3 and ClustalX, algorithm development and computational exercises were performed in MATLAB with the aid of a Statistical Toolbox extension module whenever applicable (MathWorks, Natick, Mass., USA).

Analysis of Sequence Similarity

Alignments were performed with FASTA3 [30, 31], using various substitution matrices in combination with different gap penalty settings. Several members of the BLOSUM [32] and PAM [33, 34] matrix families included in the FASTA3 program were evaluated. Moreover, 83 different substitution matrices obtained from GenomeNet [35] were probed for relevance to protein allergen classification

and 25 of them were tested on identical data sets. The highest ranked matrices were more closely evaluated. They were BLOSUM 50, BLOSUM 80, MDM 20, Protein Identity and PHAT-T75-B73 (a transmembrane matrix) [36] (http://blocks.fhcrc.org/~pauline/PHAT_matrices/). Alignment tests were conducted with default gap penalties, which in FASTA3 are -12/-2 for gap opening penalty/gap extension penalty, as well as very high penalties (-10⁶-10⁶).

Sequence Representation Using Alignment Scores and Alignment Length

Descriptors for training and testing of the classifiers evaluated here were generated from pairwise local alignments, based on FASTA3 [31]. Each sequence was aligned against a set of 120 preselected prototype allergen sequences. The default features employed for classification were the alignment score and alignment length of the best (highest alignment score value) matching alignment.

Organization of the Data Sets

The 1,007 non-allergen records were randomly sampled (without substitution) to obtain a negative (non-allergen) test set containing 272 sequences and a negative design (training) set of 128 sequences. Analogously, the allergen sequence data set was randomly divided into a positive example (allergen) test set of 70 amino acid sequences and a master set containing 248 amino acid sequences. The master set was further partitioned into a prototype sequence set of 120 sequences and a positive design (training) set of 128 sequences (fig. 1A). To furnish maximal amino acid sequence diversity in the prototype set, the model set sequences were hierarchically arranged in 120 different clusters, and one example (sequence) from each cluster was randomly selected to the prototype set. The clustering procedure used in this step involved the creation of fixed length numerical vectors, based on physicochemical features [37].

Supervised Classifiers Used to Predict Allergenicity

kNN Classifier. A kNN classifier is fed with a set of design (training) patterns from each class (the learning process). Subsequently, each test pattern is classified by identification of the *k* closest (most similar) neighbors (using for example the standard Euclidean distance). The determination of the class prediction depends on the majority class belonging among the *k* closest neighbors. In this paper, *k* was equal to 9 since preliminary results favored this parameter setting (data not shown), which is also in line with our earlier results [18].

Bayesian Linear and Quadratic Gaussian Classifiers. The Bayes rule defines the posterior class probability as:

$$P(c_i|x) = \frac{P(c_i)p(x|c_i)}{p(x)}$$

where *c_i* denotes class *i* and *x* denotes the feature vector (pattern). The class conditional probability density functions (pdfs) *p(x|c_i)* can be estimated based on design (training) examples. For a new test pattern a Bayesian classifier will then assign the pattern to the class with the highest probability. In the linear Gaussian classifier, the pdfs *p(x|c_i)* *i* = 1,2, are both assumed to be multivariate Gaussian with the same covariance matrix yielding a classifier with a hyperplane decision boundary which in the special case of only two features reduces to a straight line. In the quadratic Gaussian classifier, both pdfs are also assumed to be Gaussian but the covariance matrices are assumed to be different, yielding quadratic decision boundaries which, for the special case of only two features, reduce to either a single ellipse or a pair of (separated) hyperbolic curves.

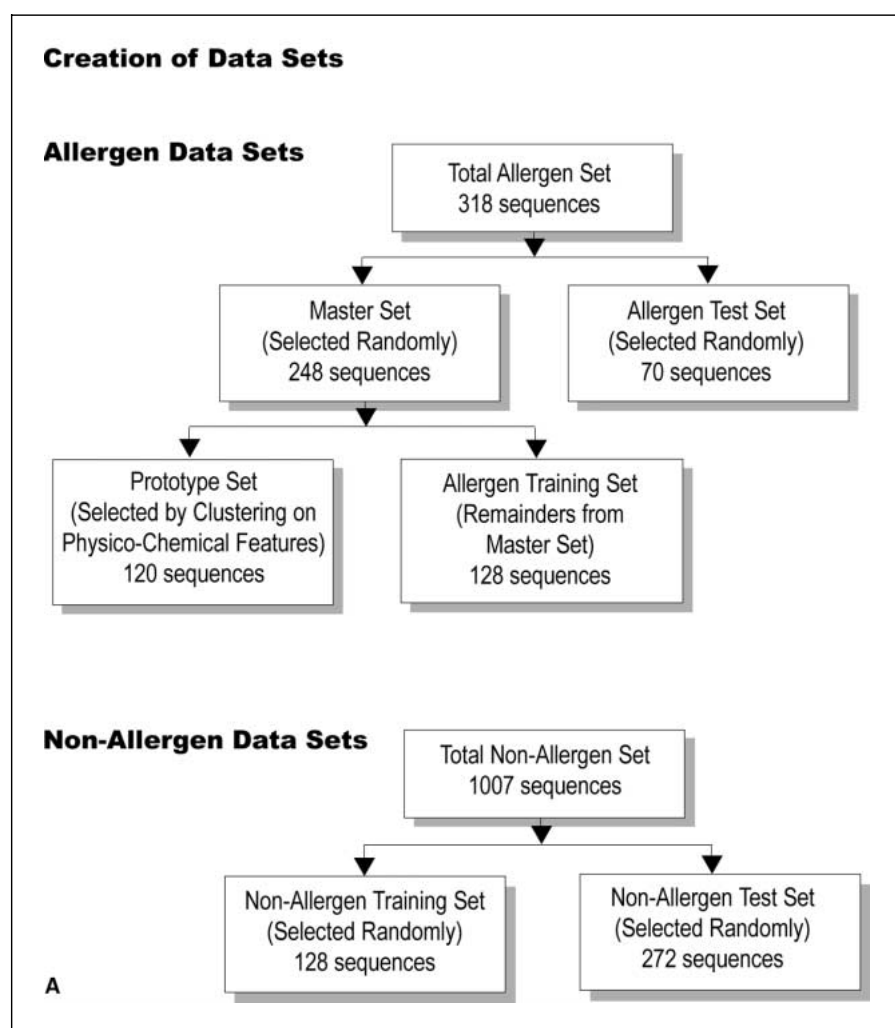


Fig. 1. Flow chart depicting the process for each of the 200 tests of classification performance. **A** The data sets used in each test were created from the total allergen and non-allergen amino acid sequence set, respectively.

Model Validation

Alternative to ROC Curves. A ROC curve depicts sensitivity (true-positive fraction) on the ordinate versus the fraction of false-positive scores ($1 - \text{specificity}$) on the abscissa, i.e. the trade-off between sensitivity and specificity for a diagnostic test [21]. In the context presented here, the ordinate represents probability of allergen detection and the abscissa describes probability of erroneous classification of non-allergens as allergens. A ROC plot is generated by incremental threshold changes. As an alternative to ROC curves, each performance plot presented here encompasses four distinct curves: detection rate and false alarm level, each expressed as mean rates and 95% confidence limits.

Cross Validation by Means of Repeated Random Holdout. A variant of the k -fold cross validation method was used. The data set was repeatedly randomly divided into test and training sets; more specifically a randomly selected holdout test was created k different times. By creating numerous holdout tests (in this project $k = 200$), mean values of the classification result (probability of detection and probability of false alarms), including 95% confidence boundaries, were obtained for the accordingly tested classifiers.

Results

A Novel High-Quality Database

The first main result of the work reported here is the novel high-quality database obtained with 318 allergen sequences and 1,007 sequences carefully selected to be non-allergens. The importance of high-quality data both for design and evaluation of any classification scheme cannot be overemphasized and therefore a substantial effort was put into this work. The quality is reflected in many ways, for example only allergens occurring in at least two out of five publicly available databases were selected. Moreover, the new database is largely devoid of isoallergens. This is because large numbers of such records may introduce a biased representation, resulting in the enhancement of features not necessarily associated with allergenic properties. Similarly, records annotated as

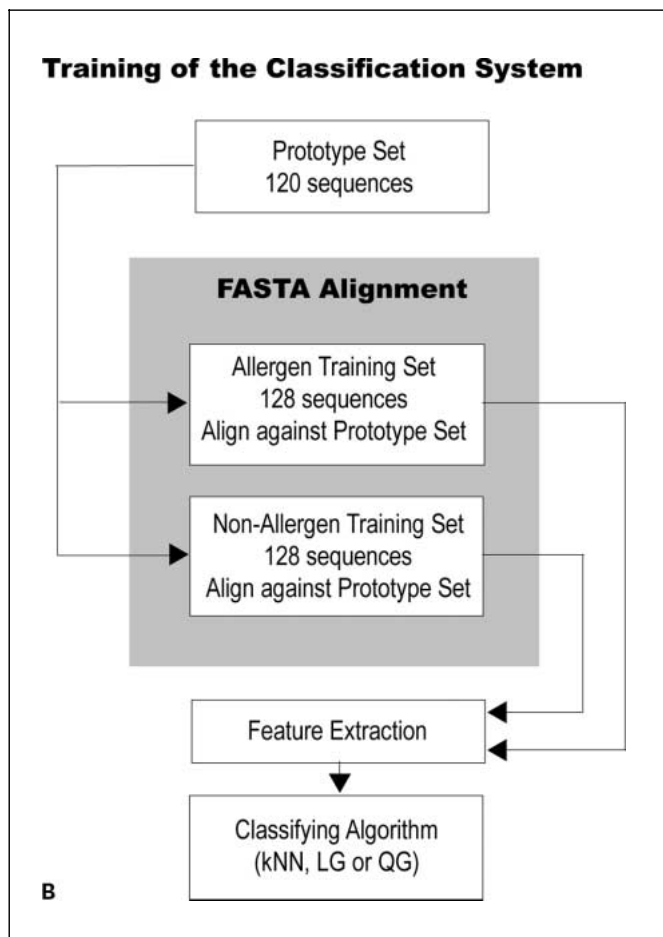
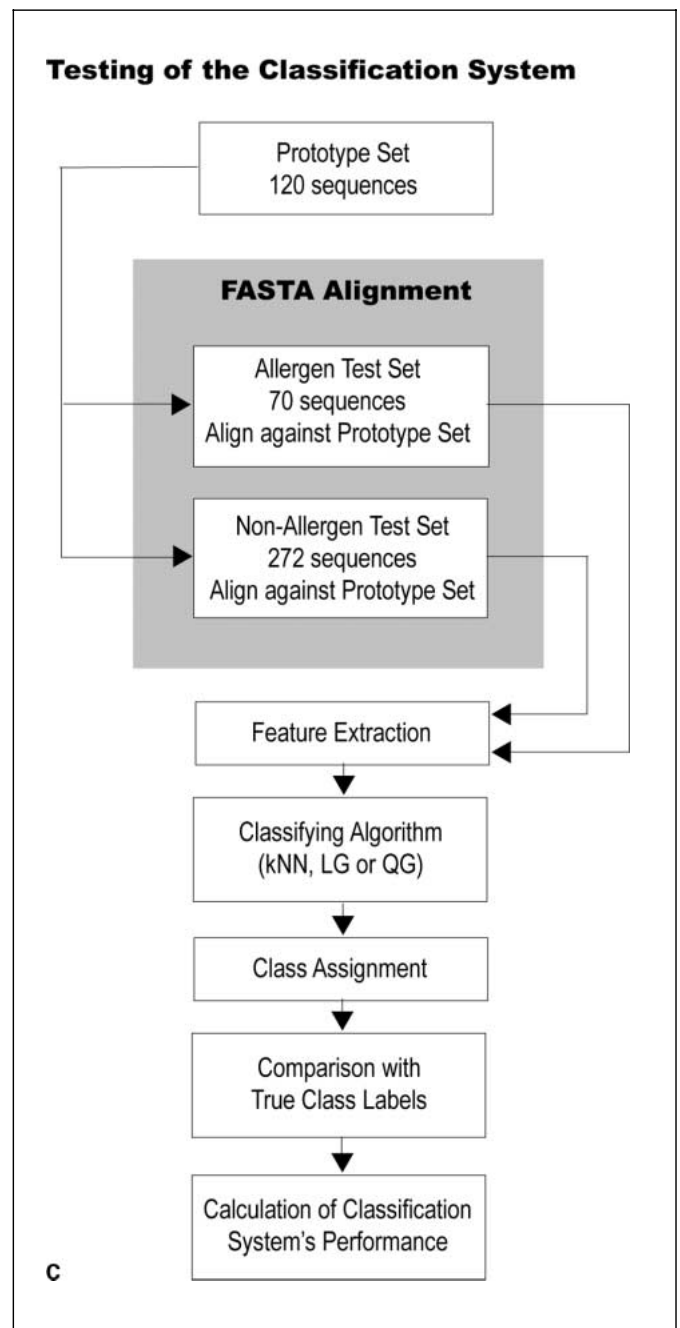


Fig. 1. B The output from the alignment between the training sets and the prototype set was used to instruct either of three classifying algorithms (kNN, LG or QG), to discriminate between allergenic and non-allergenic amino acid sequences. **C** The designed classifier was fed with the output from the alignment between the test sets and the prototype set. For each test sequence the designed classifier predicted the class and this prediction was compared with the true class identity of the sequence in order to calculate the misclassification rate. In the FASTA local alignment procedure, five different substitution matrices were tested in combination with two gap penalty settings.

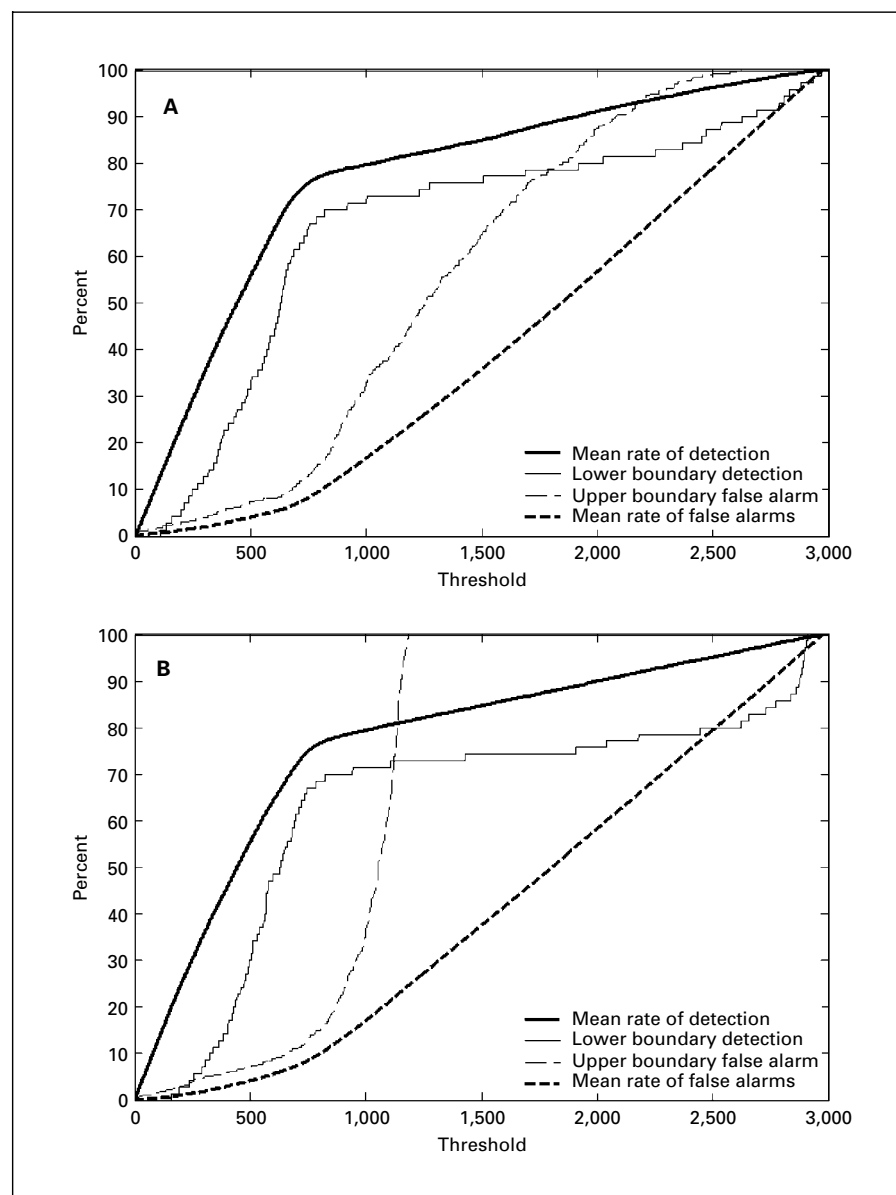


'fragments' (with a sequence shorter than 200 amino acids) were also omitted to reduce the possibility of introducing entries devoid of allergen-entailed motifs.

In the process of extracting the non-allergens, 3,000 sequences, selected for not being entailed to allergy, were compiled into a local repository. To minimize the probability of introducing allergenic proteins into this set, two search criteria were applied. Firstly, only organisms con-

taining identified protein allergens were mined, because poorly documented commodities and other food sources are conceivably at higher risk of expressing hitherto undetected allergens. Secondly, records showing the text strings 'allergen', 'atopy' or derivatives thereof, were omitted.

Fig. 2. Plots showing the trade-off between the rate of allergen detection versus the rate of false alarms for four different classification models after 200 holdout tests. Besides the mean values of the two aforementioned measures (bold lines), the boundaries (with 95% confidence) showing the minimal rate of allergen detection and the maximal rate of false alarms, respectively, are displayed (thin lines). The indices of the threshold values used to generate the curves are listed on the abscissa. **A** Linear Gaussian classifier using alignment score and alignment length (BLOSUM 80 with default gap penalties) as feature vector. **B** Quadratic Gaussian classifier using alignment score and alignment length (BLOSUM 80 with default gap penalties) as feature vector.



Low-Dimensional Visualization of the Allergen Distribution

In order to visually explore the distribution of allergens, fixed-length numerical descriptors of the amino acid sequences based on physicochemical properties [38, 39], were obtained. Then the accordingly translated sequences were subjected to hierarchical clustering. No allergen-specific clustering was, however, disclosed in such tests (not shown). Analogously, visualization based on principal component analysis did not reveal any difference between the two sequence classes (not shown). These findings are in compliance with those reported earlier, based on small-

er sets of allergenic/non-allergenic protein sequences [18].

Statistical Validation of Classifier Systems

Applied Methods for Validation of Classifier Algorithms, Substitution Matrices and Gap Penalty Settings. In this paper, three distinct classification algorithms were extensively tested, each in combination with five substitution matrices and two gap penalty settings. Accordingly, 30 different classification models were compared with regard to performance. The two key questions of interest for a particular classifier system including its feature vec-

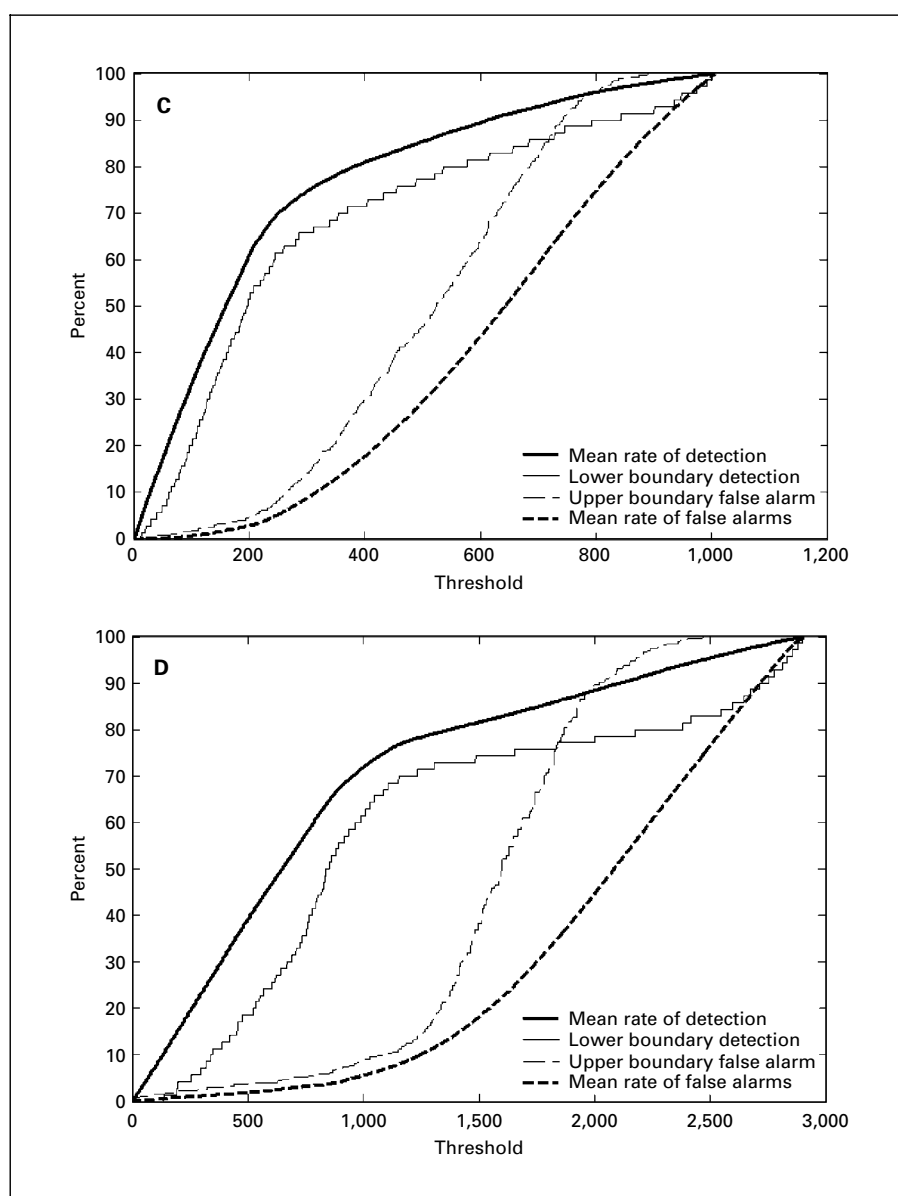


Fig. 2. C kNN classifier using alignment score and alignment length (BLOSUM 50 with default gap penalties) as feature vector. **D** Linear Gaussian classifier using combined alignment scores (BLOSUM 80 with default gap penalties and protein identity matrix with default gap penalties) as feature vector.

tor composition are: (1) What is the average performance one could expect to observe? (2) How large is the variability (variance) around this average?

As indicated earlier, approximate answers to the two questions above were obtained by means of 200 re-samplings of a carefully assembled database consisting of 318 allergens and 1,007 non-allergens in total. In each re-sampling, 128 non-allergens (negative design set) and 248 allergens (120 in the prototype set and 128 in the positive design set) were randomly selected and used to design the classification model (fig. 1B). The remaining 70 allergens and 272 randomly selected non-allergens were collected

into a test set used to estimate the performance of the classifier system designed (fig. 1C).

Statistically Meaningful Alternative to ROC Curves. Conventional ROC analysis assumes that the estimates of the sensitivity and specificity values used to draw the ROC curves are almost exact, i.e. without significant estimation errors. However, in most real world applications, the number of test examples available for estimation are relatively few which results in uncertain estimates. In practice this means that different ROC curves obtained by means of small sample sets from the same problem may vary significantly. In this work, each classifier system

Table 1. Maximal rate of false alarms (PFA) given the desired minimal rate of allergen detection (PD) for the four different classification models illustrated in figure 2 (the average values – mean PD and mean PFA – are also included)

Classifier	95% confidence		Mean PD, %	Mean PFA, %
	minimal PD, %	maximal PFA, %		
Linear Gauss	66	13	75	8
	70	16	77	10
	76	51	82	27
	80	83	90	53
Quadratic Gauss	66	12	74	8
	70	17	77	11
	76	100	89	54
	80	100	95	77
kNN (k = 9)	66	13	73	8
	70	22	78	13
	76	40	83	24
	80	52	87	34
Linear Gauss (combined alignment scores as feature vector)	66	10	74	6
	70	11	77	8
	76	57	83	25
	80	95	91	56

was redesigned and tested 200 times to obtain information about how the performance (and therefore the ROC curves) may vary from one design and test to the next. In order to distinctively show the variability, a completely new sort of performance curve was devised and implemented (fig. 2A–D). For each classifier the performance, measured by 200 randomly selected holdout tests, is described by four different curves: the mean rate of detection, the mean rate of false alarm, the lower boundary for the rate of detection (with 95% confidence) and the upper boundary for the rate of false alarm (with 95% confidence). Hence, these plots demonstrate the classifiers' average performance as well as a measure of the variance. The indices of the threshold values used to generate the curves are listed on the abscissa, and the performance is given in percent on the ordinate. For a given threshold the lowest detection limit as well as the highest false alarm limit for the classifier are shown (both calculated with 95% confidence).

Validation of Classification Results. The best results, for each classifier, are shown by means of the new performance curves introduced here (fig. 2). At a modest to reasonably high detection level of allergens, using alignment score and alignment length as features, no drastic differ-

ence between the classification algorithms was seen. Nonetheless, the LG outperformed the others by a narrow margin to QG at the 70% level of detection (table 1). This recognition rate appears as being the most useful trade-off level between allergen detection and erroneous classification of sequences devoid of association with allergy. The QG algorithm became very error prone at higher detection rates, whereas the kNN proved comparatively robust as it resisted a steep increase in false alarm numbers at excessively tuned settings (table 1, fig. 2A–C). Furthermore, while the optimal scoring matrix for LG and QG classifiers were BLOSUM 80, the best results for kNN were achieved with BLOSUM 50. Notably, default gap opening and extension penalties consistently outperformed the high penalty settings. Our judgment is that the obtained data, based on the aforementioned features, give an overall preference for the LG classifier.

Although a merged vector of the two above-described FASTA3 features has proven useful, we also combined alignment scores from different scoring matrices into a single vector, which ultimately fed each of the three classification algorithms. Using an accordingly composed feature vector, the false-alarm-type accuracy of the LG algorithms improved markedly revealing an error rate of only 11% at the 70% detection level (both values at 95% confidence limits; table 1, fig. 2D). This setting conferred no improvement in performance, however, for the remaining two classification algorithms, in relation to earlier results (not shown).

Allergenicity Test for 2S Albumin in Brazil Nut and Cry9C in Genetically Modified Corn. Thresholds for generating ROC curves were optimized for each of the classification systems described above and depicted by performance curves in figure 2A–D (LG, QG, kNN and LG with combined alignment scores). A detection level of 70% or more in conjunction with a minimal false alarm rate was judged as optimal. Accordingly, this setting was chosen for specific tests on the Brazil nut 2S albumin and the modified Cry9C protein [40–42]. The query procedure included the 2S albumin amino acid sequence in the test data set in 44 of the 200 sampling cycles. The readout was as follows: All classification models assigned the 2S albumin as an allergen in all tests, except for the QG model, which returned a single misclassification in 44 cycles. The modified Cry9C was not included in any of the training or prototype sets used to create the classification models and was therefore tested in all of the 200 cycles. This protein was consistently classified as a non-allergenic protein, except by the kNN model, which returned an allergen assignment in 131 of the 200 sampling cycles.

Discussion

In the first exploratory step of the work reported here, no apparent physicochemical features amongst protein allergens were revealed using low-dimensional visualization methods, such as principal component analysis and hierarchical clustering. This is in line with our earlier results on smaller data sets and also with other reports on structural comparison between proteins associated with allergenic potential and those devoid of such features [13, 18].

Local alignment algorithms are designed to find sequence regions with high similarity, whereas global alignment searches for overall sequence resemblance. As argued by Gendel [9, 10], the former sort of algorithm can be assumed to have a preference in this context, because allergenicity preferably involves short sequences. The applicability of this approach to allergen identification was originally shown for a relatively small set of amino acid sequences. Recently, we have demonstrated the suitability of local alignment for the identification of potential food allergens, using a much larger set of amino acid sequences [18]. The applicability of supervised classification algorithms to the identification of evolutionary relatedness or functional similarity has been widely demonstrated [43–46]. In the above-mentioned paper on the specific classification of sensitizing food allergens versus non-allergens, selected output from local alignment was allowed to successfully train a kNN learning system [18].

The work reported here extends earlier work on the prediction of allergenicity in many directions. First of all, results from the use of supervised classification algorithms applied to alignment features for the prediction of general IgE-mediated allergenicity has never been reported before. Secondly, although food allergenicity prediction using a similar approach has been described elsewhere [18], the work reported here is a significant extension with respect to the number of classifiers used, the number of different alignment procedures considered, the size and the quality of the databases used, and the level of the statistical evaluation performed.

The course to designing classifiers for the prediction of allergenicity and not only food protein allergenicity was chosen for at least three reasons: (1) to enable a larger set of positive examples (allergens), (2) to tackle the problem of cross-reactivity, either between sensitizing airborne protein allergens and food proteins or vice versa, and (3) a more general predictor for allergenicity would have a wider application range.

As regards safety assessment on allergenic potential of transgene proteins of GM crops there are, however, con-

flicting reports on whether a match of 6, 7 or 8 consecutively occurring identical amino acids to known allergens should trigger an alarm for allergenic potential [6, 9, 10, 12]. Although not fully validated in the context of protein-allergen-associated motif coverage or the frequency of erroneous hits, using any of the aforementioned settings, extension procedures have been suggested to improve on the identity match approach. In 2002, Gendel [10] suggested a two-step assessment protocol involving identity scoring followed by similarity scoring, the latter based on an evolutionary matrix. Kleter and Pfeijnenburg [12], on the other hand, recommended identity search based on the 6-amino-acid match concept, prior to comparison of accordingly targeted matches with known IgE-binding epitopes and a subsequent test for antigenicity likelihood. In addition to the classification methods presented in this paper, we have also performed preliminary tests designed to disclose the applicability of matches of six contiguous identical amino acids to predict allergenicity. Results from such tests revealed an incorrect classification of over 50% of the 1,007 presumed non-allergenic sequences (data not shown). This finding gives support to recent reports demonstrating that exact matches of six or fewer amino acids are likely to occur frequently by chance [11, 12].

Although local alignment features have so far outperformed global alternatives for the purpose of signalling protein allergenicity, epitopes that are brought together by protein folding are not easily identified using this approach. Such discontinuous epitopes may be scattered over large regions and engage only few residues on some spots. Moreover, it is widely known that peptide conformation can vary with the folding of the host protein. In this perspective, local alignment may appear as a blunt tool for the generation of descriptors to machine learning systems aimed at the classification of protein allergens. Still, the statistical evaluation performed in the work reported here convincingly demonstrates a statistically significant classification efficacy of the highest ranked systems. To obtain this, however, four provisions were identified as being prominent:

1. Modest Gap Extension and Opening Penalties in the Alignments. The regions accordingly recognized should still be considerably shorter than those encompassing discontinuous epitopes. Our findings on a preference for modest (default) gap opening/extension penalties, however, are open to speculations on the occurrence of several adjacent epitopes, possibly of mixed types, in the high-score- and/or high-length-aligned regions. Moreover, an

implication of discontinuous peptide motifs across relatively short regions, which do not necessarily belong to the epitope category, cannot be excluded in this context. For example, motifs that confer or contribute to resistance to degradation by low pH, proteases and heat are likely to occur more commonly in sensitizing food proteins, compared to other sorts of protein allergens.

2. *A High-Quality Repository of Non-Allergens.* A high-quality repository of not only allergens but also non-allergens was required for both a successful classifier design and, even more importantly, statistical evaluation of the performance.

3. *Integration of a Set of Suitable Supervised Classification Algorithms.* With a view to the great number of possible classification algorithms available, a suitable subset had to be selected. Three important issues guided the selection: The classifiers should be (1) computationally fast, (2) simple to interpret and (3) have a documented history of being relatively robust in a wide range of applications.

4. *Tuning to a Sensible Trade-Off between Detection and False Alarm.* The classifier systems used must be tuned to conduct the classifications at a sensible trade-off between detection (sensitivity) and false alarm. The conducted statistical evaluation was accordingly depicted in the form of novel performance curves instead of conventional ROC curves.

The performance of each of the presented models was extensively validated using 200 sampling cycles. This approach enabled us to show model performance as regards classification accuracy over a range of settings. Additionally, the Brazil nut 2S albumin and the modified Cry9C amino acid sequences were used as query examples in separate tests. We found the former to be consistently assigned as an allergen, whereas the latter was returned as a non-allergen by all models except for that based on the kNN algorithm. Although the potential allergenicity of the StarLink Cry9C is still debated, there are several indications on non (type I hypersensitivity)-offensive properties, especially as regards the oral exposure route [47, 48].

A commonly overlooked problem in statistical validation of designed classifier systems is the random variation obtained due to small test sets. In other words, the variance and thus the confidence intervals obtained and presented here do not only reflect variations due to small design sets but also variations due to the small size of the test sets. A careful analysis of the relative impact of these contributions is still a research topic far beyond the scope

of this work, but one should keep in mind for example that the use of the $70 + 272 = 342$ test examples here only guarantees that the length of the 95% confidence interval of the performance error estimate obtained is less than 10% [20]. Therefore, although the performance curves presented are based on a proper statistical procedure, they should not be overinterpreted. More specifically, the estimated variance behind the 95% confidence intervals shown in the new performance curves presented depend (nontrivially) on the variance caused by the relatively small number of design examples as well as on the variance caused by the relatively small number of test examples used.

While being immensely useful for research on evolutionary relatedness, the efficacy of FASTA programs drops drastically to identify sequence resemblance below 20–35% [49, 50]. Several strategies for improved relationship identification have been described, including multiple sequence alignments, elaborated formulas for p values of gapped local alignments and adaptive sequence alignments based on Bayesian algorithms [43, 51–55]. Given an alignment score between two sequences, the p value is the chance that an alignment of two randomly selected sequences would obtain this score or higher. Notably, multiple sequence approaches are reported to perform considerably better than standard local alignment procedures [56]. Some of these extensions to the FASTA program may have a potential for conferring improvements also to the classification course described here. Nonetheless, it is important to emphasize a significant difference between the course described in this article and searches for evolutionary relatedness, being the vastly predominant application for alignment programs and also a basis for entailed modifications of such methods. According to current knowledge, sequence motifs essential to the offending function of protein allergens are far from equivalent to domains with sequence similarities that occur within evolutionary protein families. Hence, implementations, which enhance the detection of evolutionary protein family members, may not automatically improve the classification of amino acid sequences associated with allergy. Furthermore, whether to use alignment scores or p values for the evaluation of sequence similarity depends on the problem at hand. The aims of this study included an incentive to identify allergen-specific motifs present in both long and short amino acid sequences, and since p values are length dependent, extended sequences encompassing short motifs could principally be overlooked. Accordingly, alignment scores were preferred in our work.

The alignment score and alignment length features are strongly correlated, which might suggest that one of them might be superfluous. We have earlier shown, however, that the use of both features is justified since a single feature proved insufficient for optimal classification of food allergens in conjunction with the kNN algorithm [18]. Nevertheless, the preliminary results described in this paper, using a feature vector that combines the alignment scores from two different substitution matrices, indicate a need for additional feature testing.

A classification model, based on the MEME motif discovery tool, has been reported recently [17]. The data on classification performance are promising, but the figures are not easily compared with those of our models. This is largely due to differences in example sequences used for test purposes. The set on allergen amino acid sequences employed by Stadler and Stadler [17] is about twofold larger compared with that of our repository, and dissimilarities with respect to allergen isoform representation may contribute to the differences observed between the respective classification approaches. Moreover, the two different non-allergen sequence sets, presented in the aforementioned report, encompass a set of synthetic amino acid sequences generated by randomization of the allergens and a second set consisting of SwissProt sequences longer than 25 amino acids not included in the allergen set. In this paper, however, 1,007 protein sequences, carefully selected from SWall to be devoid of association with allergenicity, were used to create a non-allergenic data set.

Ideally, an assessment procedure designed with the purpose to specifically detect allergenic proteins should of

course identify all such sequences, whereas other sequences should be left unmarked. However, as reported here, this seems to require more powerful features than the alignment score and the alignment length. Intriguingly, alignment score only, although extracted from two separate scoring matrices, produced the overall best classification accuracy at useful trade-off levels between levels of detection and false alarm. Albeit the aforementioned setting conferred improvement to only one of the three tested classification algorithms, the result suggests expanded testing of combinations of scoring matrices and FASTA3 outputs. Some additional courses to improvement are expansion of our current local bioinformatic repository of allergenic amino acid sequences and the establishment of a yet more rigorously quality-assured set of non-allergenic amino acid sequences. The demonstrated statistical variability across the tested classification algorithms, scoring matrices and gap opening/extension penalties obtained should be useful in the design of an integrated assessment scheme for potential protein allergenicity but also as a reference for future improvements in the particular bioinformatic learning system approach evaluated here.

Acknowledgments

Mrs. Maj Olausson is acknowledged for expert assistance on computer graphics. The Swedish Agency for Innovation Systems (VINNOVA), the Swedish National Board for Laboratory Animals (CFN) and the Faculty of Science and Technology (Uppsala University) supported this work. We thank Dr. René Crevel (Unilever Research, Bedford, UK) for critically reading the manuscript.

References

- 1 Young E, Stoneham MD, Petrukevitch A, Barton J, Rona R: A population study of food intolerance. *Lancet* 1994;343:1127–1130.
- 2 Scientific Committee on Plants: The occurrence of severe food allergies in the EU: Report of experts participating in task 7.2. Scientific Cooperation Program in the EU, 1998.
- 3 Sampson HA: Food allergy. 1. Immunopathogenesis and clinical disorders. *J Allergy Clin Immunol* 1999;103:717–728.
- 4 Robinson C: Genetic Modification Technology and Food. Consumer Health and Safety. ILSI Europe Concise Monographs. Brussels, ILSI Europe, 2001.
- 5 Metcalfe DD, Astwood JD, Townsend R, Sampson HA, Taylor SL, Fuchs RL: Assessment of the allergenic potential of foods derived from genetically engineered crop plants. *Crit Rev Food Sci Nutr* 1996;36(suppl):S165–S186.
- 6 FAO/WHO: Evaluation of allergenicity of genetically modified foods. Joint FAO/WHO Expert Consultation on Allergenicity of Foods Derived from Biotechnology. Rome, 2001.
- 7 EU (2003) Regulation (EC) No 1829/2003 of the European parliament and of the council of 22 September 2003 on genetically modified food and feed. *Off J Eur Commun* L268: 1–23. http://europa.eu.int/eur-lex/pri/en/oj/dat/2003/L_268/L_26820031018en00010023.pdf.
- 8 Lack G, Chapman M, Kalsheker N, King V, Robinson C, Venables K: Report on the potential allergenicity of genetically modified organisms and their products. *Clin Exp Allergy* 2002; 32:1131–1143.
- 9 Gendel SM: The use of amino acid sequence alignments to assess potential allergenicity of proteins used in genetically modified foods. *Adv Food Nutr Res* 1998;42:45–62.
- 10 Gendel SM: Sequence analysis for assessing potential allergenicity. *Ann NY Acad Sci* 2002; 964:87–98.

- 11 Hileman RE, Silvanovich A, Goodman RE, Rice EA, Holleschak G, Astwood JD, Hefle SL: Bioinformatic methods for allergenicity assessment using a comprehensive allergen database. *Int Arch Allergy Immunol* 2002;128:280–291.
- 12 Kleter GA, Peijnenburg AA: 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* 2002;2:8.
- 13 Aalberse RC: Structural biology of allergens. *J Allergy Clin Immunol* 2000;106:228–238.
- 14 Ayuso R, Lehrer SB, Reese G: Identification of continuous, allergenic regions of the major shrimp allergen pen a 1 (tropomyosin). *Int Arch Allergy Immunol* 2002;127:27–37.
- 15 Ivanciuc O, Schein CH, Braun W: Data mining of sequences and 3D structures of allergenic proteins. *Bioinformatics* 2002;18:1358–1364.
- 16 Bailey TL, Elkan C: Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc Int Conf Intell Syst Mol Biol* 1994;2:28–36.
- 17 Stadler MB, Stadler BM: Allergenicity prediction by protein sequence. *FASEB J* 2003;17:1141–1143.
- 18 Zorzet A, Gustafsson M, Hammerling U: Prediction of food protein allergenicity: A bioinformatic learning systems approach. *In Silico Biol* 2002;2:525–534.
- 19 Hastie T, Tibshirani R, Friedman J: *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. New York, Springer, 2001.
- 20 Webb A: *Statistical Pattern Recognition*, ed 2. Chichester, Wiley, 2002.
- 21 Vining DJ, Gladish GW: Receiver operating characteristic curves: A basic understanding. *Radiographics* 1992;12:1147–1154.
- 22 Allergen nomenclature. <http://www.allergen.org>.
- 23 Farrp Allergen Database. <http://www.allergenonline.com>.
- 24 The Allergen Database. <http://www.csl.gov.uk/allergen>.
- 25 The Allergen Sequence Database (Food). <http://www.iit.edu/~sgendel/foodallr.htm>.
- 26 The Protall Database. <http://www.ifr.bbsrc.ac.uk/protall>.
- 27 Boeckmann B, Bairoch A, Apweiler R, et al: The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res* 2003;31:365–370.
- 28 SRS at ExPASy. <http://us.expasy.org/srs5>.
- 29 Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG: The CLUSTAL_X windows interface: Flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res* 1997;25:4876–4882.
- 30 Pearson WR, Lipman DJ: Improved tools for biological sequence comparison. *Proc Natl Acad Sci USA* 1988;85:2444–2448.
- 31 Pearson WR: Flexible sequence similarity searching with the FASTA3 program package. *Methods Mol Biol* 2000;132:185–219.
- 32 Henikoff S, Henikoff JG: Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci USA* 1992;89:10915–10919.
- 33 Dayhoff MO, Schwartz RM, Orcutt BC: A model of evolutionary change in proteins. Matrices for detecting distant relationships; in Dayhoff MO (ed): *Atlas of Protein Sequence and Structure*. Washington, National Biomedical Research Foundation, 1978, vol 5, pp 345–358.
- 34 Schwartz RM, Dayhoff MO: Matrices for detecting distant relationships; in Dayhoff MO (ed): *Atlas of Protein Sequence and Structure*. Washington, National Biomedical Research Foundation, 1979, vol 5, pp 353–358.
- 35 Substitution Matrix Deposit. <http://www.genome.ad.jp/dbget/aaindex.html>.
- 36 Ng PC, Henikoff JG, Henikoff S: PHAT: A transmembrane-specific substitution matrix. Predicted hydrophobic and transmembrane. *Bioinformatics* 2000;16:760–766.
- 37 Li W, Jaroszewski L, Godzik A: Clustering of highly homologous sequences to reduce the size of large protein databases. *Bioinformatics* 2001;17:282–283.
- 38 Sandberg M, Eriksson L, Jonsson J, Sjostrom M, Wold S: New chemical descriptors relevant for the design of biologically active peptides. A multivariate characterization of 87 amino acids. *J Med Chem* 1998;41:2481–2491.
- 39 Venkatarajan MS, Braun W: New quantitative descriptors of amino acids based on multidimensional scaling of a large number of physical-chemical properties. *J Mol Model* 2001;7:445–453.
- 40 Gander ES, Holmstroem KO, De Paiva GR, De Castro LA, Carneiro M, Grossi de Sa MF: Isolation, characterization and expression of a gene coding for a 2S albumin from *Bertholletia excelsa* (Brazil nut). *Plant Mol Biol* 1991;16:437–448.
- 41 Lambert B, Buysse L, Decock C, et al: A *Bacillus thuringiensis* insecticidal crystal protein with a high activity against members of the family Noctuidae. *Appl Environ Microbiol* 1996;62:80–86.
- 42 CBH-351. <http://www.agbios.com/dbase.php?action=ShowProd&data=CBH-351&format=LONG>.
- 43 Krogh A, Brown M, Mian IS, Sjolander K, Haussler D: Hidden Markov models in computational biology. Applications to protein modeling. *J Mol Biol* 1994;235:1501–1531.
- 44 Nielsen H, Brunak S, von Heijne G: Machine learning approaches for the prediction of signal peptides and other protein sorting signals. *Protein Eng* 1999;12:3–9.
- 45 Lin K, May AC, Taylor WR: Amino acid substitution matrices from an artificial neural network model. *J Comput Biol* 2001;8:471–481.
- 46 Karwath A, King RD: Homology induction: The use of machine learning to improve sequence similarity searches. *BMC Bioinformatics* 2002;3:11.
- 47 Teshima R, Watanabe T, Okunuki H, et al: Effect of subchronic feeding of genetically modified corn (CBH351) on immune system in BN rats and B10A mice. *Shokuhin Eiseigaku Zasshi* 2002;43:273–279.
- 48 Bernstein JA, Bernstein IL, Bucchini L, et al: Clinical and laboratory investigation of allergy to genetically modified foods. *Environ Health Perspect* 2003;111:1114–1121.
- 49 Goldstein RA, Luthey-Schulten ZA, Wolynes PG: Optimal protein-folding codes from spin-glass theory. *Proc Natl Acad Sci USA* 1992;89:4918–4922.
- 50 Rost B: Twilight zone of protein sequence alignments. *Protein Eng* 1999;12:85–94.
- 51 Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res* 1997;25:3389–3402.
- 52 Park J, Teichmann SA, Hubbard T, Chothia C: Intermediate sequences increase the detection of homology between sequences. *J Mol Biol* 1997;273:349–354.
- 53 Bailey TL, Gribskov M: Methods and statistics for combining motif match scores. *J Comput Biol* 1998;5:211–221.
- 54 Bailey TL, Gribskov M: Estimating and evaluating the statistics of gapped local-alignment scores. *J Comput Biol* 2002;9:575–593.
- 55 Zhu J, Liu JS, Lawrence CE: Bayesian adaptive sequence alignment algorithms. *Bioinformatics* 1998;14:25–39.
- 56 Park J, Karplus K, Barrett C, Hughey R, Haussler D, Hubbard T, Chothia C: Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods. *J Mol Biol* 1998;284:1201–1210.