

Prediction of food protein allergenicity: a bioinformatic learning systems approach  
Anna Zorzet, Mats Gustafsson<sup>1</sup> and Ulf Hammerling<sup>2</sup>, \*

Division of Toxicology, National Food Administration, P.O. Box 622, SE-751 26 Uppsala, Sweden  
<sup>1</sup> Signal and Systems Group, Uppsala University, P.O. Box 528, SE-751 20 Uppsala, Sweden  
<sup>2</sup> Phone: +46 18 17 57 52; Fax: +46 18 17 14 33; E-mail: ulfh@slv.se

\* corresponding author

Edited by E. Wingender; received September 5, 2002; revised and accepted December 19, 2002;  
published January 07, 2003

## Abstract

Food hypersensitivity is constantly increasing in Western societies with a prevalence of about 1-2% in Europe and in the USA. Among children, the incidence is even higher. Because of the introduction of foods derived from genetically modified crops on the marketplace, the scientific community, regulatory bodies and international associations have intensified discussions on risk assessment procedures to identify potential food allergenicity of the newly introduced proteins.

In this work, we present a novel biocomputational methodology for the classification of amino acid sequences with regard to food allergenicity and non-allergenicity. This method relies on a computerised learning system trained using selected excerpts of amino acid sequences. One example of such a successful learning system is presented which consists of feature extraction from sequence alignments performed with the FASTA3 algorithm (employing the BLOSUM50 substitution matrix) combined with the k-Nearest-Neighbour (kNN) classification algorithm. Briefly, the two features extracted are the alignment score and the alignment length and the kNN algorithm assigns the pair of extracted features from an unknown sequence to the prevalent class among its k nearest neighbours in the training (prototype) set available.

91 food allergens from several specialised public repositories of food allergy and the SWALL database were identified, pre-processed, and stored, yielding one of the most extensively characterised repositories of allergenic sequences known today. All allergenic sequences were classified using a standard one-leave-out cross validation procedure yielding about 81% correctly classified allergens and the classification of 367 non-allergens in an independent test set resulted in about 98% correct classifications.

The biocomputational approach presented should be regarded as a significant extension and refinement of earlier attempts suggested for in silico food safety assessment. Our results show that the framework described here is powerful enough to become useful as part of a multiple-procedure

test scheme that also depicts other evaluation approaches such as solid phase immunoassay and tests for stability to digestions.

Key words: food allergy, risk assessment, computational toxicology

## Introduction

Food hypersensitivity is constantly increasing in Western societies with a prevalence of about 1-2% in Europe and in the USA, and among children the incidence is even higher [Bruijnzeel-Koomen et al., 1995; Sampson, 1999]. Notably, the incidences of food allergenic-induced anaphylactic reactions have increased significantly during later years [Moneret-Vautrin and Kanny, 1995; Sampson, 2000]. Although more than 90% of allergic food hypersensitivity are accounted for by eight food groups, of which peanuts, soybeans, tree nuts and crustaceans comprise the most commonly involved sources, at least 160 foods are reported to show association with sporadic allergic reactions [FAO, 1995; Hefle et al., 1996]. Many food allergens are glycoproteins with a molecular weight ranging from 10 kD to 70 kD. Moreover, resistance to proteolysis by low pH and digestive enzymes as well as relatively good heat stability, are other typical features of such proteins [Astwood et al., 1996; Fuchs and Astwood, 1996]. Nonetheless, food allergens may be devoid of most such features whereas proteins with no history of causing allergy may show one or several of the aforementioned properties. Throughout the years, the *in vitro* and *in vivo* test batteries for allergenic potential, such as specific and targeted serum screening, protein resistance to digestive treatment at low pH, the skin prick test and various animal models has gradually improved [Wal, 1999]. Nonetheless, no single universal and reliable experimental assay *in vitro*, *in vivo* or *in silico* has hitherto been reported for the assessment of allergenic products.

Especially for the safety assessment of genetically modified (GM) crops and GM foods, allergenic potential of transgene products has emerged as an increasingly critical issue, in discussions within the regulatory framework. A particular alarming example of the necessity of safety assessment for GM foods is a genetic engineered soybean crop, which expressed a methionine/cystein-rich protein (2S albumin) from Brazil nuts. No clear offending capacity of the novel food was detected in rat test models, but subsequently immunochemistry experiments revealed the presence of a major allergen in the GM soybean [Melo et al., 1994; Nordlee et al., 1996].

A systematic stepwise approach to the evaluation of food allergenicity of transgene-encoded proteins in engineered crops, commonly referred to as the IFBC/ILSI decision tree, was suggested in 1996 and appeared in a slightly revised form in the next year [Metcalf et al., 1996; Taylor, 1997]. This strategy involves several inspection methods for allergic potential of a protein, and is based on the assumption that the overall predictability of allergenicity will be enhanced by a combination of tests. Since that time, this overall evaluation scheme has undergone several slight modifications and a substantially revised version of the original process was recently presented by FAO/WHO [FAO, 2001]. Amino acid sequence alignment of a candidate transgene to target sequences of known protein allergens represents a key test procedure in the aforementioned decision tree. Regardless of the source of transgene, with respect to known food allergenicity potential, inspections for sequence homology is mandatory in the scheme and a positive output will attribute the test protein a likely allergenic property.

In the present work, we have addressed the issue of bioinformatics in predictive allergology and present a refined procedure, relative to published reports, for the screening for allergenic potential of food proteins. A biocomputational scheme comprises a fundamental part of the presented

assessment approach. Although not mandatory to the procedure itself, validated in-house excerpts from public databases of amino acid sequences have embodied key prerequisites for system construction. The method presented here is suggested to be used in the context of multi-procedure risk assessment outlines, such as that reported by FAO/WHO [FAO, 2001; Taylor, 2002].

The described procedure is based on a computerised learning system trained using selected excerpts from amino acid sequence alignments. The general methodology is illustrated by an example system consisting of a feature extraction part based on the FASTA3 sequence alignment algorithm and a classification part based on the k-Nearest-Neighbour (kNN) pattern classifier. The system is trained to classify food allergens and non-food allergens into distinct categories and is extensively validated *in silico* for both of the aforementioned protein categories. Leave-one out tests [Hastie et al., 2001; Stone, 1974] are used for the validation of the allergens due to the scarcity of available sequences, and a separate set of amino acid sequences are used for validation of the non-allergens. The system allows for convenient examinations and almost immediate data readouts using a standard 1.1 GHz microprocessor. It is applicable to the analysis of potential food allergenicity of GM crop proteins but should be useful for test of any type of food protein.

## Materials and methods

### Database mining and establishment of in-house repositories

Five separate repositories of allergen amino acid sequences were manually inspected for amino acid sequences of food allergens [Allergen nomenclature; FARRP Allergen Database; The Allergen Database; The Allergen Sequence Database; The Protall Database]. Records that appeared in at least two repositories, and with a satisfactory annotation and an association number from one or more among the well-curated databases, were merged into an in-house assembly. Additionally, a search was conducted in SWALL (SwissProt and TrEMBL) [Bairoch and Apweiler, 2000], with the aid of an ExPASy Molecular Biology server at the Swiss Institute of Bioinformatics [ExPASy Molecular Biology Server], using "allergen" as a text string. Items of this group that qualified as food allergens were deposited into the aforementioned assembly. In total 91 records were allocated to the food allergen training set (see also paragraph below).

A cohort of 200 non-allergenic sequences (negative training data set) was created by extraction from the SWALL repository, using the following search criteria: Organism: *Lycopersicon* (tomato), *Apium* (celery) or *Pyrus* (pear). Moreover, three clusters of exclusion criteria were applied: i) "allergen" or "allergy" (all text); ii) "lipid-transfer protein", "cupin", "chitinase", "profilin" (all text) [Aalberse, 2000; Breiteneder and Ebner, 2000]; iii) fragment (all text), 0:30 (sequence length). Most entries originate from tomato. Finally, a test data set, NDSTest (negative test data set) containing 367 non-allergenic amino acid sequences, was established. This amino acid sequence pool was established in analogy with the design of the two aforementioned training data sets, except that 75 of the records were obtained from the following organisms: *Malus* (apple), *Gadus* (cod) and *Prunus* (peach, cherry and apricot). The reason for using the organisms mentioned above to compose the negative training set and the test set is that they only contain one or a few reported allergens. Listings of protein accession numbers of the abovementioned data records, both negative and positive training and test sets, are available at <http://www.slv.se/foodallergy/Tables>. Tables showing the false negatives (allergens) and false positives (non-allergens) can also be found at this address.

## Trimming of amino acid sequences

Based on annotation attached to imported amino acid sequences that were filed into either of the food allergenic (positive) or non-allergenic (negative) training data sets, signal peptides and/or chloroplast target peptides were manually removed whenever found in the annotations. The accordingly processed items were filed into two distinct repositories, trimmed training data set of food allergens (PDSTTraining) and trimmed training data set of non-allergens (NDSTTraining), respectively. These sets were used for training of the example learning system. The test data set (NDSTest) was not trimmed accordingly due to two factors: Firstly, the annotations of the sequences were found insufficient with regard to the potential presence of signal peptides. Secondly, and most importantly, the use of non-trimmed test sequences accurately mimics real-life conditions where it is not always possible to ascertain whether a sequence contains a signal peptide or not.

## Computer programming

Except for the FASTA3 algorithm, the mathematical computation and algorithm development were performed in MATLAB with the aid of either a Statistical Toolbox extension module or a Neural Networks Toolbox extension module where applicable (The MathWorks, Inc.).

## Analysis of sequence similarity

Alignments were conducted with the FASTA3 program, using the BLOSUM50 substitution matrix [Pearson, 1995; Pearson, 2000; Pearson and Lipman, 1988] in combination with the gap opening penalty set to -12 and the extension gap penalty set to -2 (these are default settings). A short computer program was specifically written to enable local automated pair-wise alignment tests. Features from each alignment were merged into feature vectors that in turn fed the kNN classification algorithm. The features extracted were alignment score and alignment length.

## Sequence representation using alignment scores and alignment length

The sequence representation used is based on pair-wise alignment of sequences using the FASTA3 program [Pearson, 2000]. Briefly, a sequence was aligned against all the allergen sequences in PDSTTraining and patterns (features) were extracted from the  $m$  alignments with the highest alignment score values.

## Tuning of the learning system

Very briefly, in the kNN classifier used, a set of prototype patterns from each class were stored (learned) in the computer memory. Then each test pattern was classified by identification of the  $k$  closest (most similar) neighbours (using the standard Euclidean distance) and then by determination of the majority class representation among the  $k$  closest neighbours. The two parameters to tune in this simple learning system are therefore the number  $k$  of neighbours and the  $m$  number of alignments used for feature extraction (resulting in  $2m$  elements in the pattern vector, two elements for each alignment). These two parameters were tuned using leave-one-out cross validation on the training data set.

## Classification performance

The percent correct classification, PC, was computed as

$$PC = \frac{c1 + c2}{N1 + N2} \times 100$$

c1: number of correct classifications for allergenic sequences

c2: number of correct classifications for non-allergenic sequences

N1: number of allergenic sequences

N2: number of non-allergenic sequences

## Results

### Alternative Approaches

Besides the two features used in combination with the kNN classifier described in the Materials and Methods section above, several other features and classifiers were also evaluated (data not shown). Some of the features considered were additional outputs from the FASTA3 algorithm such as the individual lengths of the aligned sequences, the percent sequence identity within alignment matches, and results using the identity matrix as the substitution matrix (which does not at all reflect evolutionary relationships). Furthermore, amino acid properties (zz-scales) were used in combination with the auto cross covariance function to obtain features similar to the ones reported by Sandberg [Sandberg et al., 1998]. Classification methods considered other than the kNN method include linear discriminant functions, multilayer perceptron neural networks and Hidden Markov Models [Baldi and Brunak, 2001].

In conclusion, the results obtained using selected combinations of these features and classifiers often matched but never outperformed the learning system presented in this work on the particular data sets used. Some combinations, however, were very successful when applied to classification of protein families with higher inter-family similarities than allergens. More substantial efforts to tune these alternative learning systems might very well result in significantly improved performance but since this work is focused on the learning systems approach as such and not on finding the best performing system, this issue is a topic for future work.

### Feature extraction

Alignment score and alignment length, obtained from FASTA3 readouts, were selected features for merging into vectors, to feed the kNN algorithm. Other features available from the FASTA3 program output such as the percentage of identity amino acid matches and the total number of amino acids in the two proteins were also evaluated as additional features but none of them yielded any significant improvement (data not shown).

## Intra- and inter-repository sequence alignments

The 91 food allergens present in the PDSTTraining assembly served as prototypes to which all sequences were aligned. Using the preferred settings, including the BLOSUM50 scoring matrix, each alignment score was plotted against the corresponding alignment length forming a two-dimensional feature point. In Figure 1, panel A, the result of aligning the 91 food allergens against themselves is presented. In panels B and C of Figure 1, the results of aligning the non-allergenic sequences in the sets PDSTTraining and PDSTTest against the 91 allergenic sequences are presented, respectively. A distinctive characteristic of the plot shown in panel A is that the allergenic sequences yield widely scattered feature points with high alignment scores on average. The non-allergenic sequences, on the other hand (Fig. 1, panels B and C), yield more compact and elongated clusters of feature points characterised by relatively short alignment lengths and relatively low alignment scores. Relatively few outliers that do not follow this general pattern are also observed.

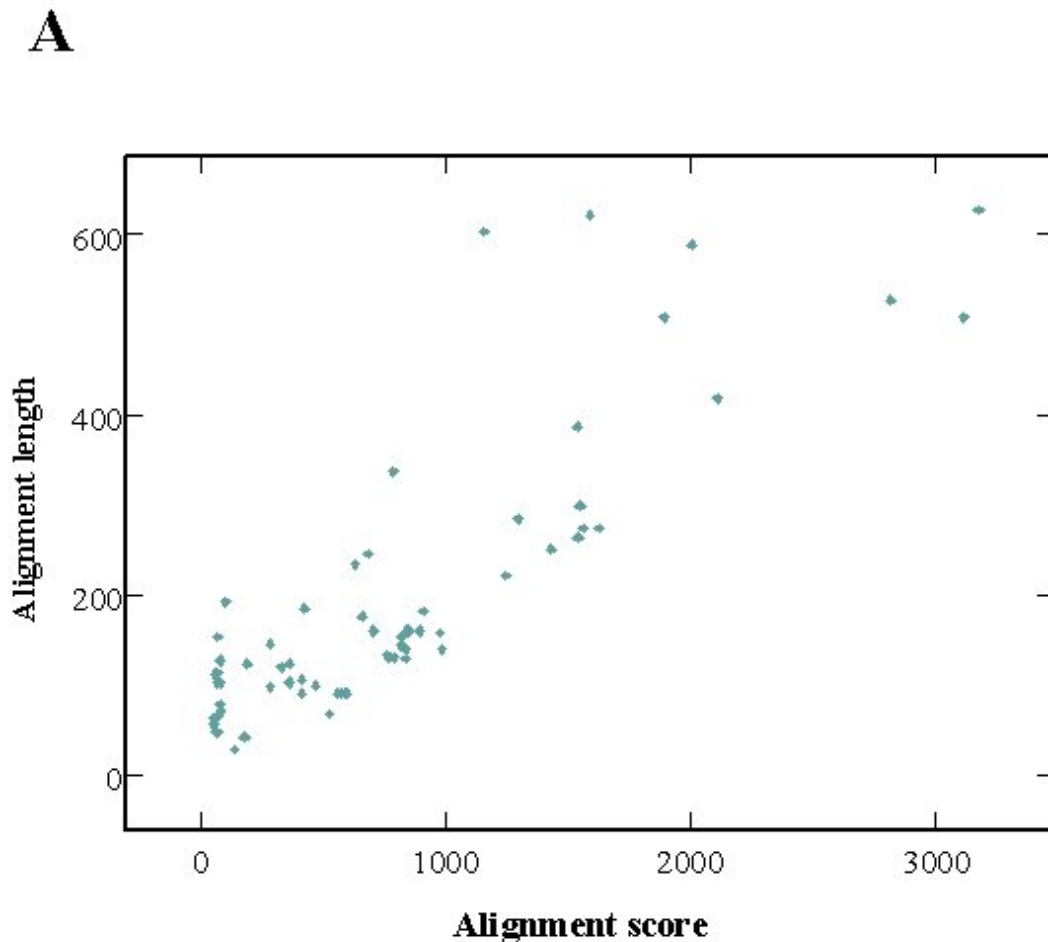


Figure 1: Raw data plots showing amino acid alignment score on the abscissa and alignment length on the ordinate. The BLOSUM50 substitution matrix was used. A. Allergen sequence alignment. The 91 food allergen sequences in PDSTTraining were aligned against themselves and the (second) best match was selected ( $m = 1$ ). B. Non-allergen sequence alignment. The non-allergens in NDSTTraining were aligned with the 91 allergens in PDSTTraining and the best match

was selected ( $m = 1$ ). C. Non-allergen sequence alignment. The non-allergens in NDSTest were aligned with the 91 allergens in PDSTTraining and the best match was selected ( $m = 1$ ).

### Adopted Approach

Regardless of parameters chosen for  $k$  and  $m$  in the learning system model, the BLOSUM50 scoring matrix significantly outperformed the identity matrix as regards the classification performance of food allergenic sequences (Figure 2, panels A and C), whereas the difference was marginal for presumed non-allergenic sequences. (Figure 2, panels B and D). Extensive evaluation of different values of the parameter  $k$  in the kNN classifier showed that the performance was not very sensitive to the exact choice of  $k$  (Figure 2, panels C and D). The combination of  $k=9$  and  $m=1$  yielded the best classification performance, about 94% correct classifications distributed as follows: Approximately 81% correct classifications of the 91 allergens using leave-one-out evaluation and about 98% correct classification of the negative test cohort (presumed non-allergens) using the 367 independent test examples (Figure 2, panels C and D). To test the ability of the system to cope with a real world problem, the now-known allergen 2S albumin from Brazil nuts [Nordlee et al., 1996] was presented to our learning system and was successfully classified as an allergen.

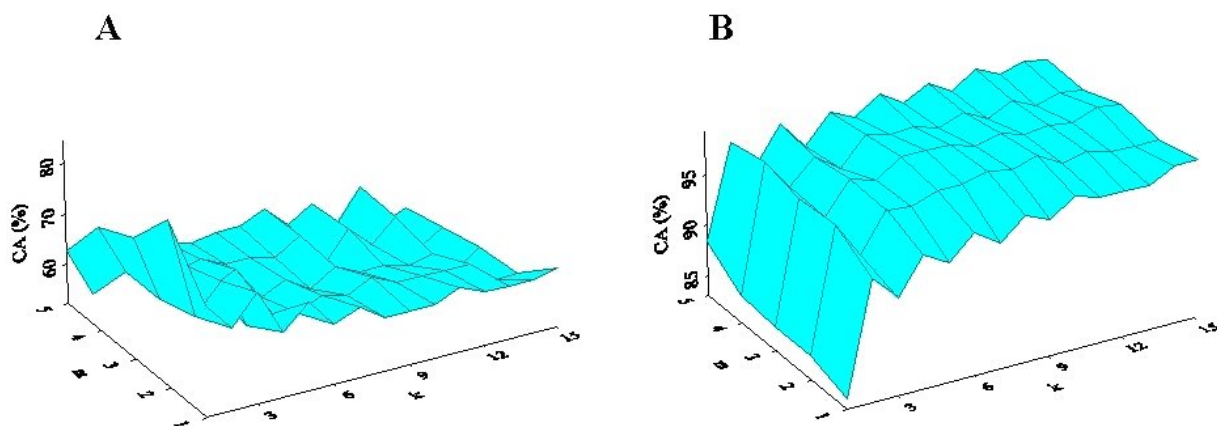


Figure 2: Surface plots showing amino acid sequence classification accuracy, expressed as the percentage correct classifications (CA (%)), as a function of the kNN-parameters  $m$  and  $k$ . Alignments were accomplished by the use of the adopted feature set, i. e. alignment score, alignment length. A. An assembly of 91 food-allergen sequences, used for training of the kNN algorithm, served as the test set. Alignments were based on the identity matrix. B. Identical to A, except that sequences selected for not having an association with allergenicity, were employed as a test material. C. Identical to A (food-allergen sequences), except that BLOSUM50 replaced the identity matrix. D. Identical to B, except for the use of BLOSUM50 as a scoring matrix.

### Discussion

Contemporary biotechnology allows for relatively precise targeted modification of the genome in plants, which can serve the purpose of inhibiting or even disrupting gene activity, activating existing genes or - more commonly - accomplish the introduction of new genes. Since the early 1990s, a large number of accordingly engineered plants varieties have been grown in field tests and

a fraction of these are currently cultivated commercially for food production. The most common genetically modified (GM) crops currently used in agriculture are soybean, corn and rapeseed expressing new proteins that impart either herbicide tolerance or insecticide resistance to the crop [IFT, 2000]. Because of the relatively high incidence of food allergenicity in Western societies, entailed to serious outcomes in many cases, special attention is given to risk assessment of the potential allergenicity of transgenic proteins expressed by GM crops. Within the frame of Codex Alimentarius, the OECD Task Force for the Safety of Novel Foods and Feeds and the EU Commission on Novel Foods, the safety of GM crops for consumers - including the issue of food allergenicity - is recognised as a high priority subject.

In the present report, we describe a bioinformatic approach for prediction of food allergenic potential of proteins, based on a trained computerised learning system that has been educated by data derived from sequence alignment matches, using defined sets of amino acid sequences. For a particular learning system described in detail, verification tests revealed approximately 81% correctly classified food allergens and approximately 98% accuracy for amino acid sequences devoid of association with allergenic potential. The purpose with this bioinformatic (and biocomputational) classification method is, though, not to disqualify more conventional evaluation practices but rather to extend and refine early bioinformatic steps in food safety assessment [Gendel, 1998; FAO, 2001]. Hence, we believe our method, in analogy with that based on simple sequence alignment tests, be best used in the context of a multiple-procedure test scheme that also depicts other evaluation approaches such as solid phase immunoassay and tests for stability to digestion.

Although representing a very low fraction of the total number, several outliers were found in alignments of presumed non-allergenic sequences (negative records) with food allergenic counterparts. Also proteins from the negative training set with remarkably high scores were considered. An intriguing finding emerged upon inspection of the function of the most prominent outliers. Out of the seven top scoring negatives (negative training and test sets), five are chlorophyll a/b binding proteins from tomato, pear and apple. If all outliers with an alignment score larger than 150 are considered, there are 12 chlorophyll a/b-binding proteins. One presumed allergen, CB23\_APIGR from celery, is also a member of this functional category. However, after close inspection of the references found for this allergen, it appears that the allergenicity of this protein can be questioned. It is, however, found in several databases as well as annotated as an allergen in Swissprot. Assuming that this protein is not in fact an allergen, it shows that our system is robust enough to handle erroneous data in the training set.

An evaluation of the applicability of the identity amino acid matrix, and biochemical and evolutionary counterparts, for the identification of food allergenic proteins, has been reported elsewhere [Gendel, 1998]. Despite that very few sequences were used as test material, this approach revealed short stretches of identity between known and presumptive food allergen sequences [Gendel, 1998]. This merit of the identity substitution matrix, which may be at advantage in manual inspection procedures, must be recognised.

In the present work, the identity matrix and BLOSUM50 were tested in our classification model. Although the difference in performance between the identity and the evolutionary matrix was almost insignificant in some tests of our provisional models, the latter still performed notably better on an overall basis and especially in conjunction with the adopted feature set for alignment (Figure 2; compare top and bottom panels). This finding suggests that alternatives to the identity matrix, including those based on biochemical/evolutionary properties of amino acids, may confer improvement on the described model's predictive accuracy. Other important issues in this context,

which need further studies, are of course the gap extension and gap opening penalties, which have profound influence on the typical length of blocks of identities in an alignment.

Concerning sequence similarity, the FAO/WHO report delineates a bifurcated query procedure for determination of sequence similarity between a test protein and a set of allergenic molecules. The first criterion is based on a specified degree of overall sequence similarity (>35%), whereas the second is oriented towards stretches of identical amino acids (6 or more). There is, though, an appreciable risk of 6 contiguous amino acids occurring by chance. Therefore, verification of cross-reactivity is warranted when only this criterion is met in the alignment procedure between a test sequence and the allergenic sequence data [FAO, 2001]. A recent report demonstrates large numbers of spurious hits using an alignment setting of 6 amino acids joined in order as a limit for alarm, and 8 amino acids is shown to produce more relevant sequence identifications [Hileman et al., 2002].

FASTA3 is used as the alignment software for several reasons. Firstly, it is recommended in the FAO/WHO protocol [FAO, 2001]. Secondly, it is widely used, well described and operates smoothly with our in-house software package. When reviewing the classification results, it is evident that the use of an alternative alignment algorithm, reporting several local alignments (e. g. BLAST), would presumably not improve the classification results. This is because even the best possible alignment (provided by FASTA3) to each such allergen sequence already yields very low score values and readouts from several local alignments would accordingly not help much.

We would like to emphasise that the approach presented here differs from the suggested FAO/WHO protocol in several respects. Firstly, both alignment score and alignment length data (according to the FASTA3 algorithm) were employed as features. This implies that sequence similarity in the form of short identical stretches and occurring as more elongated similar motifs, but without the aforementioned output criteria, is considered. Secondly, the features were merged into vectors that in turn fed a kNN classification algorithm and training was accomplished by the use of sequences of two distinct functional categories. Thirdly, we have used a selected set of only food allergen sequences as a (positive) reference repository. Concerning the set of reference allergens, the FAO/WHO protocol provides recommendations to include all allergen amino acid sequences indexed in SwissProt. We believe that the use of a more restricted set of sequences improves the classification accuracy, in particular by reducing the number of false positives. Moreover, an accordingly selected repository is likely to promote the automatic identification of characteristic features, which are unique to food allergens such as an increased likelihood of being resistant to proteolysis. The high ratings for classification of the two protein categories indicate a good recognition power of the described prediction model for several key features of food allergens.

As mentioned above, data output from both alignment score and alignment length were used to create merged vectors. The basic assumption is that certain features of the alignments should differ if the sequence in question is an allergen or a non-allergen. Also of interest in this context is that the alignment score and alignment length features are strongly correlated, which might suggest that one of them is redundant. Our use of both features, however, is justified by our experimental findings, which reveal that use of a single feature yields lower performance on the data sets used (results not shown) and from the theoretical fact that the two features are not identical and thus partly contain different information.

Finally, one should note that the food-allergenic data set is clearly a limiting resource, in particular because of the high structural diversity among such proteins [Aalberse, 2000]: certain structural classes may simply be represented by too few examples to qualify for adequate training of a learning algorithm. Since the availability of allergens sequence data is very limited, however, the

risk of such bias cannot entirely be avoided. Although at a modest pace the number of identified food allergens is constantly accumulating over time, thereby offering improved possibilities to achieve higher predictive performance using the bioinformatic learning systems methodology described here. This development (in combination with future work on e. g. alternative sequence representations and more advanced and well-tuned learning systems) offers great promise in the efforts towards substantial improvements of the results presented in this work.

## Acknowledgements

This work was supported by the Swedish Agency for Innovation Systems (VINNOVA). We are grateful to Dr. René Crevel at Unilever Research, Bedford, UK, and Dr. Clare Mills at the Institute of Food Research, Norwich, UK, for helpful information on food allergenic amino acid sequences and for rewarding discussions on other specific issues of food allergenicity.

## References

- Aalberse, R. C. (2000). Structural biology of allergens. *J. Allergy Clin. Immunol.* 106, 228-238.
- Allergen nomenclature. <http://www.allergen.org>; 2001-10-11.
- Astwood, J. D., Leach, J. N. and Fuchs, R. L. (1996). Stability of food allergens to digestion in vitro. *Nat. Biotechnol.* 14, 1269-1273.
- Bairoch, A. and Apweiler, R. (2000). The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.* 28, 45-48.
- Baldi, P. and Brunak, S. (2001). *Bioinformatics: The Machine Learning Approach*. MIT Press.
- Breiteneder, H. and Ebner, C. (2000). Molecular and biochemical classification of plant-derived food allergens. *J. Allergy Clin. Immunol.* 106, 27-36.
- Bruijnzeel-Koomen, C., Ortolani, C., Aas, K., Bindslev-Jensen, C., Bjorksten, B., Moneret-Vautrin, D. and Wuthrich, B. (1995). Adverse reactions to food. European Academy of Allergology and Clinical Immunology Subcommittee. *Allergy* 50, 623-635.
- ExPASy Molecular Biology Server of the Swiss Institute of Bioinformatics. <http://www.expasy.org> ; 2001-10-11.
- FAO. (1995). Report of the FAO Technical Consultation on Food Allergies. Food and Agriculture Organization of the United Nations, Rome.
- FAO. (2001). Evaluation of allergenicity of genetically modified foods. Joint FAO/WHO Expert Consultation on Allergenicity of Foods Derived from Biotechnology. Food and Agricultural Organization of the United Nations (FAO), Rome, Italy.
- Farrp Allergen Database. <http://www.allergenonline.com>; 2001-10-11.
- Fuchs, R. L. and Astwood, J. D. (1996). Allergenicity assessment of foods derived from genetically modified plants. *Food Technol.* 50, 83-88.
- Gendel, S. M. (1998). The use of amino acid sequence alignments to assess potential allergenicity of proteins used in genetically modified foods. *Adv. Food Nutr. Res.* 42, 45-62.
- Hastie, T., Tibshirani, R. and Friedman, J. (2001). *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. Springer-Verlag, New York.
- Hefle, S. L., Nordlee, J. A. and Taylor, S. L. (1996). Allergenic foods. *Crit. Rev. Food Sci. Nutr.* 36 Suppl., S69-S89.
- 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.
- IFT (2000). IFT Expert Report in Biotechnology and Foods: Introduction. *Food Technol.* 54, 124-136.

- Melo, V. M. M., Xavier-Filho, J., Lima, M. S. and Prouvost-Dannon, A. (1994). Allergenicity and tolerance to proteins from Brazil nut (*Bertholletia excelsa* H. B. K. ). *Food Agric. Immunol.* 6, 185-195.
- Metcalf, 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.
- Moneret-Vautrin, D. A. and Kanny, G. (1995). [Food-induced anaphylaxis. A new French multicenter study]. *Bull. Acad. Natl. Med.* 179, 161-172, 178-184; discussion 173-167.
- Nordlee, J. A., Taylor, S. L., Townsend, J. A., Thomas, L. A. and Bush, R. K. (1996). Identification of a Brazil-nut allergen in transgenic soybeans. *N. Engl. J. Med.* 334, 688-692.
- Pearson, W. R. (1995). Comparison of methods for searching protein sequence databases. *Protein Sci.* 4, 1145-1160.
- Pearson, W. R. (2000). Flexible sequence similarity searching with the FASTA3 program package. *Methods Mol. Biol.* 132, 185-219.
- Pearson, W. R. and Lipman, D. J. (1988). Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci. USA* 85, 2444-2448.
- Sampson, H. A. (1999). Food allergy. Part 1: immunopathogenesis and clinical disorders. *J. Allergy Clin. Immunol.* 103, 717-728.
- Sampson, H. A. (2000). Food anaphylaxis. *Br. Med. Bull.* 56, 925-935.
- Sandberg, M., Eriksson, L., Jonsson, J., Sjostrom, M. and Wold, S. (1998). New chemical descriptors relevant for the design of biologically active peptides. A multivariate characterization of 87 amino acids. *J. Med. Chem.* 41, 2481-2491.
- Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *J. Royal Statistical Society* 36, 111-147.
- Taylor, S. L. (1997). Assessment of the allergenicity of genetically modified foods. *Agbiotech News Information* 9, 229N-234N.
- Taylor, S. L. (2002). Protein allergenicity assessment of foods produced through agricultural biotechnology. *Annu. Rev. Pharmacol. Toxicol.* 42, 99-112.
- The Allergen Database. <http://www.csl.gov.uk/allergen>; 2001-10-11.
- The Allergen Sequence Database. <http://www.iit.edu/~sgendel/foodallr.htm>; 2001-10-11.
- The Protall Database. <http://www.ifr.bbsrc.ac.uk/protall>; 2001-10-15.
- Wal, J. M. (1999). Assessment of allergic potential of (novel) foods. *Nahrung* 43, 168-174.