

Allergenicity prediction by protein sequence

Michael B. Stadler and Beda M. Stadler

Institute of Immunology, University of Bern, Switzerland

Corresponding author: Michael B. Stadler, Institute of Immunology, Sahlihaus 2, InselspitalCH-3010 Bern, Switzerland. E-mail:michael.stadler@insel.ch

ABSTRACT

Potential allergenicity of transgenic proteins for consumption must be investigated before their introduction into the food chain. A prerequisite is sequence analysis. We have critically reviewed the performance of the current guidelines proposed by the Food and Agriculture Organization (FAO) and the World Health Organization (WHO) for allergenicity prediction based on protein sequence and show that its precision is very low. To improve prediction, we propose a new strategy based on sequence motifs identified from a new allergen database. If tested on random test sequences and known allergens, both methods are apparently very sensitive. However, the precision of our motif-based prediction (95.5%) is superior to the current method (36.6%). We conclude that the proposed motif-based prediction is a superior alternative to the current method for use in the decision-tree approach for allergenicity assessment.

Key words: genetically modified food • allergen database • allergenicity prediction • bioinformatics • sequence motifs

Allergens inducing type I allergic responses are proteins that elicit specific IgE antibodies. The allergic reaction is triggered by allergens aggregating IgE antibodies bound to high-affinity Fc receptor (FcεRI) on mast cells and basophils (1, 2). Mediators released by activated cells cause the symptoms of allergy, such as sneezing and swelling of the mucosa, characteristic for allergic rhinitis, allergic conjunctivitis, and asthma.

Despite the great number of presently identified allergenic proteins (3), it is still not known why only few and particular proteins that humans are exposed to provoke allergic reactions. Thus, a method for allergenicity prediction would be beneficial, especially in order to prevent the inadvertent generation of new allergenic food plants by agricultural biotechnology.

In 1996, a task force of the International Food Biotechnology Council (IFBC) and the Allergy and Immunology Institute of the International Life Sciences Institute (ILSI) developed a decision-tree approach for the assessment of potential allergenicity of plants produced through agricultural biotechnology (4, 5). In 2001, FAO and WHO have modified the approach in a joint expert consultation on foods derived from biotechnology. In the consultation report (accessible at <http://www.who.int/fsf/GMfood/>), guidelines have been published for the evaluation of allergenicity of genetically modified foods. Besides biological tests concerning the protein of interest, a standard method for sequence comparison has been defined. Briefly, a protein is

considered allergenic if it shares more than 35% sequence similarity (window of 80 residues) or an identity of at least six contiguous amino acids with a known allergen.

In this study, we critically review the performance of the proposed method for the prediction based on sequence similarity. We present an automated approach for the construction and update of a local allergen sequence database. Using probabilistic sequence motifs identified from this allergen database, we propose a new approach for allergenicity prediction in order to overcome the low precision of the current method.

MATERIALS AND METHODS

Databases and Software

External sequence databases downloaded and installed locally to be used for the study: Swiss-Prot (6): Release 40.0; 101,602 proteins; obtained from: <ftp://ftp.expasy.org/databases/swiss-prot>. Randomized Swiss-Prot: Sequences of release 40 shuffled in consecutive windows of 20 amino acids. trGEN (7) Human, Release 12–19–2001; 330,743 sequences, obtained from <ftp://ftp.isrec.isb-sib.ch/pub/databases/trgen/>. Swiss-Prot allergen index: Release 16-Oct-2001; 274 protein sequences; available at: <http://www.expasy.org/cgi-bin/lists?allergen.txt>. Rice (8, 9): TIGR rice gene index (OsGI); rRelease 7.0; 10,891 protein sequences, obtained from: <http://www.tigr.org/tdb/ogi>. Remote allergen lists used for allergen database construction: <http://www.allergen.org> (10), <http://www.expasy.org/cgi-bin/lists?allergen.txt> (6), <http://www.iit.edu/~sgendel> (11). Remote sequence databases used as a source for allergen sequences: GenBank (12). PIR (13): Online access at: <http://www.ncbi.nlm.nih.gov/entrez>. A local allergen database was generated by extracting all accession numbers in the published allergen lists and downloading the corresponding sequences from the public sequence databases. Subsequently, DNA sequences were translated, and sequence variants were generated according to annotation. Finally, all redundant sequences were removed, resulting in a database containing 779 allergens (February 11, 2002). Freely available software packages downloaded and installed locally to be used for the study: pftools (14), Version 2.2, obtained from <http://www.isrec.isb-sib.ch/ftp-server/pftools>. MEME (15): Version 3.0.3, obtained from <http://meme.sdsc.edu/meme/website>. FASTA (16): Version 3.4, obtained from <ftp://ftp.virginia.edu/pub/fasta>. NCBI-BLAST (17): Version 2.2.1, obtained from <ftp://ftp.ncbi.nih.gov/blast/>. Scripts generating the allergen database and controlling iterative motif discovery and allergenicity prediction were written in Perl (<http://www.perl.org>) using extensions from Bioperl (<http://www.bioperl.org>) for sequence processing and online data retrieval.

FAO/WHO allergenicity evaluation

According to the guidelines for allergenicity evaluation of foods derived from biotechnology (full report at <http://www.who.int/fsf/GMfood/>), a query protein is potentially allergenic if it either has an identity of at least six contiguous amino acids or more than 35% sequence similarity over a window of 80 amino acids when compared with a known allergen. We have written a program that compares a query protein with each allergen and rates it allergenic, if

either of the two criteria were fulfilled. The value for identity length n could be specified as a parameter to allow for more flexible testing.

For allergen prediction in the allergen database, we slightly modified the program by removing the single-query allergen sequence from the reference allergen sequence database. Without this modification, each query sequence would be contained in the allergen database, and identical subsequences of n residues could always be found (for n not greater than sequence length).

Automated iterative allergen motif discovery

Starting with all 779 sequences in the allergen database, the following steps were performed iteratively until no motif with E-value less than 0.01 could be identified: MEME (15) (*zoops* motif match mode) was used to identify the most relevant motif of 50 residues contained in the allergen sequences. The length of 50 residues was chosen to be shorter than the mean length of a protein domain in order to prevent generation of multi-domain motifs. The mean domain length of the 974'587 Pfam (18) domains identified in Swiss-Prot and TrEMBL is 135 residues, and 79.54% are longer than 50 residues (data not shown). Shorter motif length resulted in a similar number of allergen motifs with lower prediction accuracy (data not shown).

The log-odds matrix was extracted from the MEME output and converted into a generalized profile (19) with one match state for each position in the log-odds matrix. The profile was scaled on a randomized version of Swiss-Prot using pfscale (14).

The scaled profile was used to search allergens for matching sequences using a normalized score of 8.5 as threshold. This score corresponds to less than one chance match to be expected when searching whole Swiss-Prot and TrEMBL databases (roughly 700,000 sequences).

Matching allergens were removed from the allergen database, and remaining sequences were submitted to the next iteration of motif discovery.

Of 779 allergen sequences, 644 were matched by one or several of these motifs. Of the 135 sequences that did not match an allergen motif, 78 corresponded to partial allergen sequences and could therefore not be optimally aligned to an allergen motif, and the remaining 57 were assumed to represent relatively unique allergens. As we wanted the allergen motifs to represent the common characteristics of a group of related sequences, we decided not to generate potentially unrepresentative motifs for each of these 135 allergen sequences. Nevertheless, the 135 sequences were included in the allergenicity prediction (see below).

Motif-based allergenicity prediction

A query protein sequence was predicted to be allergenic, if it matched an allergen motif with a normalized score greater than 8.5. If no matching motif was found, a second analysis step was performed for allergenicity prediction: The query sequence was aligned to the allergen sequences not matching one of the allergen motifs (135 sequences). The query was rated allergenic if it either matched an allergen motif or scored better than a certain threshold (specified as a parameter) in a pair-wise sequence alignment. A length limit of 25 amino acids was applied. This corresponds to the length of the smallest known allergen protein (bee venom melittin, 26

residues, Swiss-Prot P01501). This length limit was applied in order to exclude short fragment sequences contained in sequence databases. Where indicated, the length limit was also applied for the FAO/WHO prediction method for better comparability of both methods. Sequence alignments were performed by using *blastp* from BLAST (17), BLOSUM50 (20) similarity matrix and 12/2 as gap open/extension penalties. BLAST E-values were as discriminative between true and false positives as PRSS (16) *P* values by using 3,000 reshufflings, and both performed better than raw scores or log length normalized raw scores (data not shown). For evaluation of prediction, a test database was constructed containing true allergen sequences and three randomized version of each allergen sequence (reversed, shuffled, 20 residue window-shuffled). The resulting test database contained 2,976 protein sequences and 25% true allergens. Performance of prediction was determined by calculating *precision* and *recall* (=sensitivity) defined as:

$$\text{precision} = \text{true positives} / (\text{true positives} + \text{false positives})$$

$$\text{recall} = \text{true positives} / (\text{true positives} + \text{false negatives})$$

Ten-fold cross validation experiment

The cross-validation experiment was performed by randomly splitting the allergen database into ten parts containing equal number of sequences. The sequences contained in each part were submitted to allergenicity prediction by both FAO/WHO and motif-based methods, whereas the remaining nine parts served as allergen reference database and as source for allergen motifs. Performance was measured as *precision* and *recall*, using the allergen sequences as true positives and three randomized versions of each allergen sequence as true negatives (reversed, shuffled, 20 amino acid window-shuffled).

RESULTS

Allergen sequence database

A database of allergen sequences that is as complete as possible represents a prerequisite for bioinformatic analysis of allergens, such as defining common allergen motifs or classifying new proteins according to their similarity with known allergens. Although most sequences of allergenic proteins are known and publicly available, no single database exists that contains all of these sequences (11).

Several lists of sequence accession numbers have been published corresponding to allergen genes or proteins (10, 11). Allergens contained in the Swiss-Prot protein database (6) are collected in a dedicated index. Thus, a complete allergen database was generated by extracting all accession numbers in the published allergen lists and downloading the corresponding sequences from the public sequence databases [Swiss-Prot (6), PIR (13) and GenBank (12)]. We have written a script automatically performing this task that allows frequent database updates and facilitates the error-prone and time-consuming process of downloading the sequences manually. The allergen database used in this study was generated on February 11, 2002, and

contained 779 non-redundant protein sequences, including translated allergen genes and generated sequence variants.

Evaluation of current allergenicity prediction

It is not known whether the current method for evaluation of allergenicity proposed by FAO/WHO has been tested concerning its recall and precision. We have therefore implemented the proposed method in a program and performed allergenicity prediction for a number of different databases as described in the experimental protocol. [Table 1](#) shows the percentages of proteins predicted to be allergenic for 35% identical residues and different values of the parameter n . Using a value of 6 for the identity length n as proposed by FAO/WHO, 98.6% of the allergens in our database were correctly predicted. However, 67.3% of all proteins in Swiss-Prot were also rated as allergens, and this figure is reduced by only 0.08% if known allergens are removed from Swiss-Prot before analysis; 75.9% allergenic proteins were found in rice, and 42.9% of human trGEN sequences (7) (an automatically translated version of the human genome) were predicted to be allergenic.

For the prediction shown in [Table 1](#), signal peptides were not removed from sequences as recommended by the FAO/WHO guidelines. The reason for this simplification was that only for a minority of analyzed proteins (5.6% of Swiss-Prot proteins, and none of the proteins in rice or trGEN databases), experimental evidence on the signal peptide was available in database annotation. We therefore did not truncate sequences for allergenicity prediction. Nevertheless, we studied whether cleaving of signal peptides might influence the prediction (*Table 1*, Swiss-Prot-SP). We found the numbers of predicted allergens slightly decreased if signal peptides were removed (from 67.3 to 66.3%).

Next, we investigated the influence of increasing identity length on predicted allergens in Swiss-Prot, rice, human trGEN and allergen databases ([Table 1](#)). Augmenting the value of n drastically reduced numbers of matching n -mers and thus put more importance on the similarity criterion (35% over 80 residues) of the prediction algorithm (data not shown). This resulted in a higher stringency obtained for prediction, even though the numbers of predicted allergens were still higher than the expected percentage of real allergens (~0.4% for Swiss-Prot based on Swiss-Prot allergen index). We therefore tried to find a new approach to quantify potential cross-reactivity of a query sequence with a known allergen.

Automated iterative motif discovery in allergen database

To assess variability contained in the allergen database and to generate a minimal set of sequence motifs representing allergens, an automated iterative motif discovery was performed. Only 52 statistically relevant allergen motifs were identified in the allergen database, indicating limited variability of allergen motifs in comparison to the total number of allergens contained in the database. Of 779 allergen sequences, 644 were matched by one or several of these motifs. Of the remaining sequences, 78 corresponded to short fragment allergen sequences that could therefore not be optimally aligned to an allergen motif. Thus, the 52 allergen motifs can match over 90% of allergens longer than 50 residues. [Table 2](#) shows statistical motif qualities expressed as MEME E-values (15) for the 20 first-identified allergen motifs. E-values are an estimate of the

number of similar motifs to be expected by chance, with smaller values corresponding to more relevant motifs. If motif discovery was performed on a randomized version of our allergen database, the E-value of the best motif was $2.6 \cdot 10^{-12}$ (data not shown). The most frequent motif in the allergen database matched 101 proteins all belonging to the Bet v 1 family. This result can be explained by a bias in available allergen sequences toward well characterized birch pollen allergen Bet v 1 and related allergens, as well as by the high number of Bet v 1 isoforms. Der p 1, the clinically relevant major allergen from house dust mite, resides in the group of 16 proteins matching allergen motif 12 (AM00012). Four of the 20 allergen motifs shown in [Table 2](#) (AM00004, AM00008, AM00017, and AM00020) could not be related to a known protein family. This finding emphasizes the necessity to use allergen-derived motifs for allergenicity prediction instead of using predefined protein family signatures as those in PROSITE or InterPro (21, 22).

Motif-based allergenicity prediction

The allergen motifs identified by iterative motif discovery were used to predict potential allergenicity of query protein sequences. The allergen motifs represent a collection of the sequence families present in currently known allergens. By scanning a query sequence with these motifs, its relatedness and thus its potential allergenicity can be estimated.

The 52 allergen motifs could not match the 135 sequences. Of these, 78 corresponded to partial allergen sequences and were significantly shorter than other allergens (data not shown). The remaining 57 sequences did not have closely related sequences in the allergen database and were therefore not represented in the allergen motif collection. To correctly predict also these sequences, we designed a two-step approach for allergenicity prediction. In the first step, query sequences are compared with allergen motifs. In the second step, query sequences are aligned to the 135 unique allergen sequences (not matching an allergen motif). A similarity identified in either of both steps indicates a potential cross-reactivity of the query sequence with a known allergen.

We first studied accuracy of the approach in a ten-fold cross validation experiment ([Table 3](#)). For this experiment, the allergen database has been split into 10 random parts of equal size. Allergenicity prediction was performed for sequences in each part, using the other nine parts as allergen reference database. This approach allowed estimation of prediction accuracy for so-far unknown allergens. Non-allergen sequences were generated by randomization of true allergens. In the cross-validation experiment, the FAO/WHO method (both for $n=6$ and $n=8$) proved more sensitive than the motif-based prediction (recall of 97.0% and 92.2% vs. 86.2%, [Table 3](#)). The high recalls obtained by both methods point out the limited variability of the allergen database; even if 10% of the sequences are removed, most can still be correctly classified as allergens. A wider divergence between allergenicity prediction methods was observed in measurements of precision. Whereas the motif-based method was highly accurate (precision of 94.8%, [Table 3](#)), the FAO/WHO method reached a precision of only 37.6%. Increasing the identity length parameter n of the FAO/WHO method from six to eight amino acids improved the precision to 68%. However, further increment of n did not result higher precision (data not shown).

Using a test database containing 2'976 protein sequences and 25% true allergens, we addressed the accuracy of allergenicity prediction methods. Non-allergen sequences in the test database have been generated by randomization of allergen sequences. [Fig. 1](#) shows precision and recall of the motif-based allergenicity prediction and the prediction according to FAO/WHO guidelines by using various parameter values. Maximal precision and recall obtained by the motif-based prediction were superior to the ones obtained by the FAO/WHO method. Using a BLAST E-value cut-off of 10^{-8} (indicated by vertical line, [Fig. 1A](#)), the motif-based prediction reached a precision of 95.5% with a recall of 100%, whereas an identity length n of six amino acids for the FAO/WHO method (vertical line, [Fig. 1B](#)) resulted in a low precision of 36.6% with a recall of only 99.7%.

Finally, we directly compared motif-based and FAO/WHO prediction methods when applied to real protein sequences ([Table 4](#)). For all proteins contained in Swiss-Prot, allergenicity was predicted. As already shown in [Table 1](#), more than two-thirds of the query proteins are predicted allergenic by the FAO/WHO method. Compared with this, motif-based prediction detects only 4% allergens in Swiss-Prot, and if allergen motifs are used exclusively for prediction ([Table 4](#), motifs only method), this value is reduced further to 2.6%. To distinguish known allergens from false positives and potentially new allergens, we checked whether their sequence was contained in the allergen reference database. About 1 in 10 potential allergens predicted by the motif based methods was a true allergen, whereas only ~1 in 200 potential allergens was a true allergen when using the FAO/WHO method ([Table 4](#)).

DISCUSSION

Although the scientific community agrees on including sequence similarity in evaluation of allergenicity of foods derived from biotechnology (4, 5, 23, 24), no consensus has been reached on how to perform similarity testing (5). The aim of our study was to analyze allergen prediction methods on the basis of data acquired from known allergens and a large number of different proteins. To our knowledge, no such data-driven analysis has been performed so far. Considering the results we obtained for our reference allergen database and other general databases, we could quantify accuracy of allergen prediction. In addition, we could test a new approach for allergenicity prediction and could quantitatively compare it with current methods.

It must be pointed out that currently it cannot be claimed that a protein without sequence similarity to any known allergen might never cause an allergic reaction. Nevertheless, allergenicity prediction based on protein sequence provides an important tool to identify potential cross-reactivity with known allergens, indicating the requirement for further investigation by other techniques (4, 25).

Allergen sequence database

As allergens do not share common structural characteristics (26, 27), and epitopes recognized by the immune system cannot be predicted based on sequence data, the use of sequence similarity in allergenicity evaluation is highly dependent on a database of allergens that serves as reference. Instead of manually constructing an allergen database by literature review and database searching, we relied on previously published and regularly updated allergen lists (3, 6, 10). Our

focus was to obtain a database that would be as comprehensive as possible and to overcome the shortcomings of currently existing databases (11). Special attention was paid to sequence variants: In Swiss-Prot, variants are not contained in the database as separate sequences, but only as annotation accompanying the principal sequence entry. Processing of variant information yielded an additional 99 sequences that would otherwise not have been included in the allergen database. It will be important to update the allergen reference database on a regular basis as new allergens are identified in order to improve the performance of allergenicity prediction.

Evaluation of current allergenicity prediction

Based on current knowledge, it is not justified to consider each protein with six contiguous amino acids in common with a known allergen as potentially allergenic. This criterion predicts the majority of Swiss-Prot or rice proteins and more than 40% of human proteins as allergens, which does not reflect the numbers of true allergens to be expected in these databases. In addition, the numbers of matching 6-mers found in a protein sequence tend to increase with sequence length (data not shown). This is indicative of this prediction method producing mainly random matches whose probability increases with sequences length. FAO/WHO guidelines recommend to remove signal peptides from allergen and query sequences before allergenicity prediction. The resulting decrease of allergens predicted in Swiss-Prot may be explained by the simultaneous decrease of sequence length, and hence, a lowered probability of matching 6-mers. Prediction accuracy was not affected by signal peptide removal (data not shown).

We could show that by increasing the identity length parameter, the precision of FAO/WHO allergenicity prediction could be improved without affecting its sensitivity. However, values larger than 8 did not result in further performance gains. An adjustment of the similarity parameters (35% over 80 residues) might be necessary to optimize the performance of the approach, especially as the current values are chosen conservatively: Allergenic cross-reactivity caused by proteins sharing conformational or linear epitopes is rare at 50% identity and typically requires more than 70% amino acid identity across the full length of the proteins (26). However, we think that a motif based method is superior to a conventional sequence alignment method, as it is more flexible (see below).

Allergen motifs identified from allergen database

An improved prediction performance can thus be obtained by increasing the identity length from 6 to 8 residues and by optimizing FASTA or BLAST alignment parameters. However, local alignment search tools, such as FASTA and BLAST, exert fixed substitution scores and gap penalties, and one single similarity cut-off value would have to be defined that could discriminate between immunologically cross-reactive and non-cross-reactive proteins. Indeed, such a universal cut-off value may not exist, and individual thresholds may be necessary for different protein families. Hence, we choose to detect common sequence motifs by using profiles for allergenicity prediction that would provide us with the necessary flexibility. Profiles such as those used in our study (19) are more sensitive in detecting homologues and thus potentially cross-reactive proteins than local alignment search tools, because of their position-specific scoring system (28). In addition, each individual profile representing a motif was scaled (14)

such as the match scores produced by motif searching all become normalized and thus comparable. This is the basis for a universal threshold for immunological cross-reactivity.

Furthermore, allergen motifs serve to systematically organize allergens into groups of related and cross-reactive proteins. Our data indicate that amongst currently known allergens, there are 52 sequence families that are represented by more than one allergenic protein, and an additional group of 135 fragment sequences or unique allergens without allergenic relatives. The allergen motifs we used were identified and scaled according to an automated protocol. We have shown that allergenicity prediction based on these motifs is possible with high sensitivity and greatly improved precision compared with the current method. Nevertheless, manual inspection, such as realigning motif-containing allergens and construction of optimized profiles, has the potential to further increase prediction performance. It would be possible to include spatial information from three-dimensional structure models into the profiles, for instance to focus the profile on core residues that define the overall protein fold, and on surface-accessible residues that may be essentially determining the characteristic properties of IgE binding epitopes. Such improvements could not be realized when using local alignment search tools, as it has been proposed by FAO/WHO and by others (23). In the future, increasing numbers of allergenic proteins will be identified, resulting in a more complete set of allergen motifs and probably eliminating the need to perform pairwise alignments in motif based allergenicity prediction.

Motif-based allergenicity prediction

In a 10-fold cross validation experiment, we addressed the performance of allergenicity prediction for new allergens not contained in the allergen reference database. Both prediction methods are evidently highly sensitive, although the high recall attained by prediction according to FAO/WHO guidelines has to be ascribed to the high level of false positives produced by the method.

Precision of prediction methods was assessed by using a test set containing 25% true allergens. The non-allergenic sequences in the test set were obtained by randomization of true allergen sequences. This procedure was chosen because it alters protein fold and thus immunological properties of randomized sequences, whereas it preserves other sequence characteristics, such as compositional bias and low complexity regions, that are known to

produce statistically relevant but biologically meaningless matches (29). The results obtained for this test set ([Fig. 1](#)) are consistent with our earlier findings, namely that both methods are highly sensitive, but the FAO/WHO method produces a high number of false positives, reducing its precision. This is especially true for parameters as proposed by the FAO/WHO.

Finally, we wanted to test the prediction methods on real protein sequences. Allergenicity prediction was performed for all proteins contained in Swiss-Prot ([Table 4](#)). Only 1 in 200 potential allergens predicted according to FAO/WHO guidelines was a true allergen. It is evident that with such a high level of noise, the method cannot discriminate between non-allergens and allergens. Moreover, the method could lead to a general overestimation of allergenic potential and thus require disproportionately high efforts in clinical risk assessment. One in 12 allergens detected by the motif based method was a true allergen, and if allergenicity prediction was

performed by exclusively matching allergen motifs, 1 in 9 detected sequences was a true allergen. Indeed, the percentages of true allergens predicted by allergen motifs might be underestimated, as some of the potential proteins could actually be cross-reactive. Because of its enhanced signal-to-noise ratio, this method could be used to systematically search for new potential allergens in available sequence data.

Unfortunately, it is currently not possible to define a similarity threshold in allergenicity prediction that can truly discriminate between immunologically cross-reactive and non-cross-reactive proteins. More information on the relationship between sequence and structure and eventually adapted similarity search tools are needed for such a future estimate. Manually optimized profiles that imply information from a known three-dimensional structure may be a promising option.

Although the properties conferring allergenicity remain unknown, we have shown that our approach predicts allergenicity with good sensitivity and precision, and it performs definitely better than the current method proposed by the FAO/WHO. With growing allergen catalogs and improved methods for characterizing proteins, our approach may be further optimized and provides a reasonable tool in risk assessment to identify transgenes that require further investigation by other techniques.

ACKNOWLEDGMENTS

This work was supported by the Novartis Foundation, grant 02A07.

REFERENCES

1. Nadler, M. J., Matthews, S. A., Turner, H., and Kinet, J. P. (2000) Signal transduction by the high-affinity immunoglobulin E receptor Fc epsilon RI: coupling form to function. *Adv. Immunol.* **76**, 325–355
2. Metzger, H. (1991) The high affinity receptor for IgE on mast cells. *Clin. Exp. Allergy* **21**, 269–279
3. Gendel, S. M. (2002) Sequence analysis for assessing potential allergenicity. *Ann. N. Y. Acad. Sci.* **964**, 87–98
4. 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
5. Taylor, S. L., and Hefle, S. L. (2001) Will genetically modified foods be allergenic? *J. Allergy Clin. Immunol.* **107**, 765–771
6. Bairoch, A., and Apweiler, R. (1996) The SWISS-PROT protein sequence data bank and its new supplement TREMBL. *Nucleic Acids Res.* **24**, 21–25

7. Pagni, M., Iseli, C., Junier, T., Falquet, L., Jongeneel, V., and Bucher, P. (2001) trEST, trGEN and Hits: access to databases of predicted protein sequences. *Nucleic Acids Res.* **29**, 148–151
8. Quackenbush, J., Liang, F., Holt, I., Pertea, G., and Upton, J. (2000) The TIGR gene indices: reconstruction and representation of expressed gene sequences. *Nucleic Acids Res.* **28**, 141–145
9. Quackenbush, J., Cho, J., Lee, D., Liang, F., Holt, I., Karamycheva, S., Parvizi, B., Pertea, G., Sultana, R., and White, J. (2001) The TIGR Gene Indices: analysis of gene transcript sequences in highly sampled eukaryotic species. *Nucleic Acids Res.* **29**, 159–164
10. King, T. P., Hoffman, D., Lowenstein, H., Marsh, D. G., Platts-Mills, T. A., and Thomas, W. (1994) Allergen nomenclature. WHO/IUIS Allergen Nomenclature Subcommittee. *Int. Arch. Allergy Immunol.* **105**, 224–233
11. Gendel, S. M. (1998) Sequence databases for assessing the potential allergenicity of proteins used in transgenic foods. *Adv. Food Nutr. Res.* **42**, 63–92
12. Benton, D. (1990) Recent changes in the GenBank On-line Service. *Nucleic Acids Res.* **18**, 1517–1520
13. George, D. G., Barker, W. C., Mewes, H. W., Pfeiffer, F., and Tsugita, A. (1996) The PIR-International Protein Sequence Database. *Nucleic Acids Res.* **24**, 17–20
14. Bucher, P., Karplus, K., Moeri, N., and Hofmann, K. (1996) A flexible motif search technique based on generalized profiles. *Comput. Chem.* **20**, 3–23
15. Bailey, T. L. and Elkan, C. (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **2**, 28–36
16. Pearson, W. R. (2000) Flexible sequence similarity searching with the FASTA3 program package. *Methods Mol. Biol.* **132**, 185–219
17. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402
18. Bateman, A., Birney, E., Cerruti, L., Durbin, R., Eddy, S. R., Griffiths-Jones, S., Howe, K. L., Marshall, M., and Sonnhammer, E. L. (2002) The Pfam protein families database. *Nucleic Acids Res.* **30**, 276–280
19. Bucher, P. and Bairoch, A. (1994) A generalized profile syntax for biomolecular sequence motifs and its function in automatic sequence interpretation. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **2**, 53–61

20. Henikoff, S., and Henikoff, J. G. (1992) Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. USA*. **89**, 10915–10919
21. Falquet, L., Pagni, M., Bucher, P., Hulo, N., Sigrist, C. J., Hofmann, K., and Bairoch, A. (2002) The PROSITE database, its status in 2002. *Nucleic Acids Res.* **30**, 235–238
22. Apweiler, R., Attwood, T. K., Bairoch, A., Bateman, A., Birney, E., Biswas, M., Bucher, P., Cerutti, L., Corpet, F., Croning, M. D., et al. (2001) The InterPro database, an integrated documentation resource for protein families, domains and functional sites. *Nucleic Acids Res.* **29**, 37–40
23. 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
24. 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
25. Astwood, J. D., and Fuchs, R. L. (1996) Allergenicity of foods derived from transgenic plants. *Monogr. Allergy* **32**, 105–120
26. Aalberse, R. C. (2000) Structural biology of allergens. *J. Allergy Clin. Immunol.* **106**, 228–238
27. Bredehorst, R., and David, K. (2001) What establishes a protein as an allergen? *J. Chromatogr. B Biomed. Sci. Appl.* **756**, 33–40
28. Gribskov, M., Luthy, R., and Eisenberg, D. (1990) Profile analysis. *Methods Enzymol.* **183**, 146–159
29. Wootton, J. C., and Federhen, S. (1996) Analysis of compositionally biased regions in sequence databases. *Methods Enzymol.* **266**, 554–571

Received November 26, 2002; accepted February 27, 2003.

Table 1**Allergen prediction according to FAO/WHO guidelines.**

<i>Database</i>	<i>% allergens for a given identity length n^a</i>						
<i>Version (number of proteins)</i>	<i>n=6</i>	<i>n=7</i>	<i>n=8</i>	<i>n=9</i>	<i>n=10</i>	<i>n=11</i>	<i>n=12</i>
Allergens 02-11-2002 (779)	98.6	98.2	98.0	97.7	97.6	97.3	97.0
Swiss-Prot Release 40.0 (101'602)	67.3	17.6	8.7	7.6	7.3	7.2	7.2
Swiss-Prot w/o Allergens ^b Release 40.0 (101'328)	67.3	17.4	8.5	7.3	7.0	7.0	7.0
Swiss-Prot-SP ^c Release 40.0 (101'602)	66.3	17.1	8.6	7.5	7.2	7.2	7.2
Rice TIGR OsGI Release 7.0 (10'891)	75.9	27.6	11.3	8.0	7.3	7.2	7.2
trGEN human 12-19-2001 (330'743)	42.9	7.3	2.9	2.2	2.1	2.0	2.0

^a Query proteins were rated allergenic if either at least n consecutive residues were found in common with a known allergen, or if sequence identity with a known allergen was higher than 35 % over a window of 80 residues.

^b 274 proteins listed in Swiss-Prot allergen index were removed from Swiss-Prot.

^c Swiss-Prot sequences with known signal peptides were truncated according to annotation (FT SIGNAL) and stored in the database termed Swiss-Prot-SP.

Table 2**Allergen motifs.**

<i>Motif identifier</i>	<i>MEME E-value</i>	<i>Matching allergens</i>	<i>Predominant protein families^a</i>
AM00001	$1.8 \cdot 10^{-4123}$	101	Pathogenesis-related proteins BetvI family
AM00002	$2.0 \cdot 10^{-1477}$	68	Profilins Pollen proteins Ole e I family
AM00003	$1.3 \cdot 10^{-919}$	36	Globins
AM00004	$3.0 \cdot 10^{-845}$	35	- ^b
AM00005	$4.8 \cdot 10^{-794}$	22	SCP/Tpx-1/Ag5/PR-1/Sc7
AM00006	$2.3 \cdot 10^{-774}$	34	11-S plant seed storage proteins Caseins
AM00007	$9.2 \cdot 10^{-460}$	47	Plant lipid transfer proteins Lipases Eukaryotic thiol proteases
AM00008	$2.7 \cdot 10^{-460}$	18	- ^b
AM00009	$1.9 \cdot 10^{-323}$	14	EF-hand calcium-binding domain
AM00010	$3.5 \cdot 10^{-271}$	10	Cereal trypsin/alpha-amylase inhibitors
AM00011	$2.0 \cdot 10^{-356}$	11	Tropomyosins
AM00012	$3.2 \cdot 10^{-242}$	16	Eukaryotic thiol proteases
AM00013	$1.2 \cdot 10^{-234}$	8	Mitochondrial energy transfer proteins
AM00014	$1.6 \cdot 10^{-229}$	7	Lipocalins
AM00015	$3.3 \cdot 10^{-219}$	21	Uteroglobin family Serpins
AM00016	$4.1 \cdot 10^{-218}$	9	Caseins
AM00017	$1.2 \cdot 10^{-190}$	7	- ^b
AM00018	$7.8 \cdot 10^{-165}$	24	Plant lipid transfer proteins Chitin binding domain Barwin domain
AM00019	$2.1 \cdot 10^{-139}$	6	Enolases
AM00020	$1.3 \cdot 10^{-211}$	12	- ^b

^aPredominant protein families corresponding to allergen motifs have been identified by scanning motif containing sequences with PROSITE (21), Rel. 16.0 and updates up to Oct 2001.

^bNo predominant PROSITE protein family signature was found in matching allergens.

Table 3**Ten-fold cross validation.**

<i>Dataset^a</i>	<i>Allergen sequences</i>	<i>Motifs</i>	<i>% precision</i>			<i>% recall</i>		
			<i>FAO/WHO (n=6)</i>	<i>FAO/WHO (n=8)</i>	<i>motif based method</i>	<i>FAO/WHO (n=6)</i>	<i>FAO/WHO (n=8)</i>	<i>motif based method</i>
Set 0	75	50	36.9	68.6	97.0	97.3	93.3	86.7
Set 1	76	53	36.4	62.0	90.5	98.7	92.1	88.2
Set 2	75	49	40.0	72.5	96.9	98.7	94.7	84.0
Set 3	74	51	38.1	63.9	94.0	97.3	93.2	85.1
Set 4	77	46	37.5	67.3	98.6	97.4	93.5	88.3
Set 5	74	51	37.8	69.6	97.1	96.0	96.0	90.5
Set 6	73	49	39.4	76.2	96.8	97.3	87.7	82.2
Set 7	75	52	36.7	68.4	95.4	93.3	86.7	82.7
Set 8	70	44	36.3	66.0	87.0	98.6	94.3	85.7
Set 9	75	46	36.7	68.0	95.7	96.0	90.7	88.0
TOTAL	744	-	37.6	68.0	94.8	97.0	92.2	86.2

^aThree randomized versions of each allergen sequence in the set were generated (reversed, shuffled, 20 amino acid window-shuffled). The allergen sequences (true positives) and the randomized sequences (true negatives) were submitted to allergenicity prediction, using all other datasets as allergen reference database and a lower length limit of 25 residues.

Table 4**Comparison of FAO/WHO and motif based allergen predictions for Swiss-Prot proteins.**

<i>Prediction method</i>	<i>Potential allergens^a</i>		<i>True allergens^b</i>	
FAO/WHO	68356	67.3 %	351	0.5 %
Motif based	4093	4.0 %	351	8.6 %
Motifs only ^c	2603	2.6 %	297	11.4 %

^aPredicted allergens for Swiss-Prot proteins longer than 25 residues.

^bA potential allergen was considered a true allergen, if its sequence was contained in the allergen reference database.

^cFor the "motifs only" method, only allergen motifs were used for allergenicity prediction, without local similarity search as in step two of motif-based method.

Fig. 1

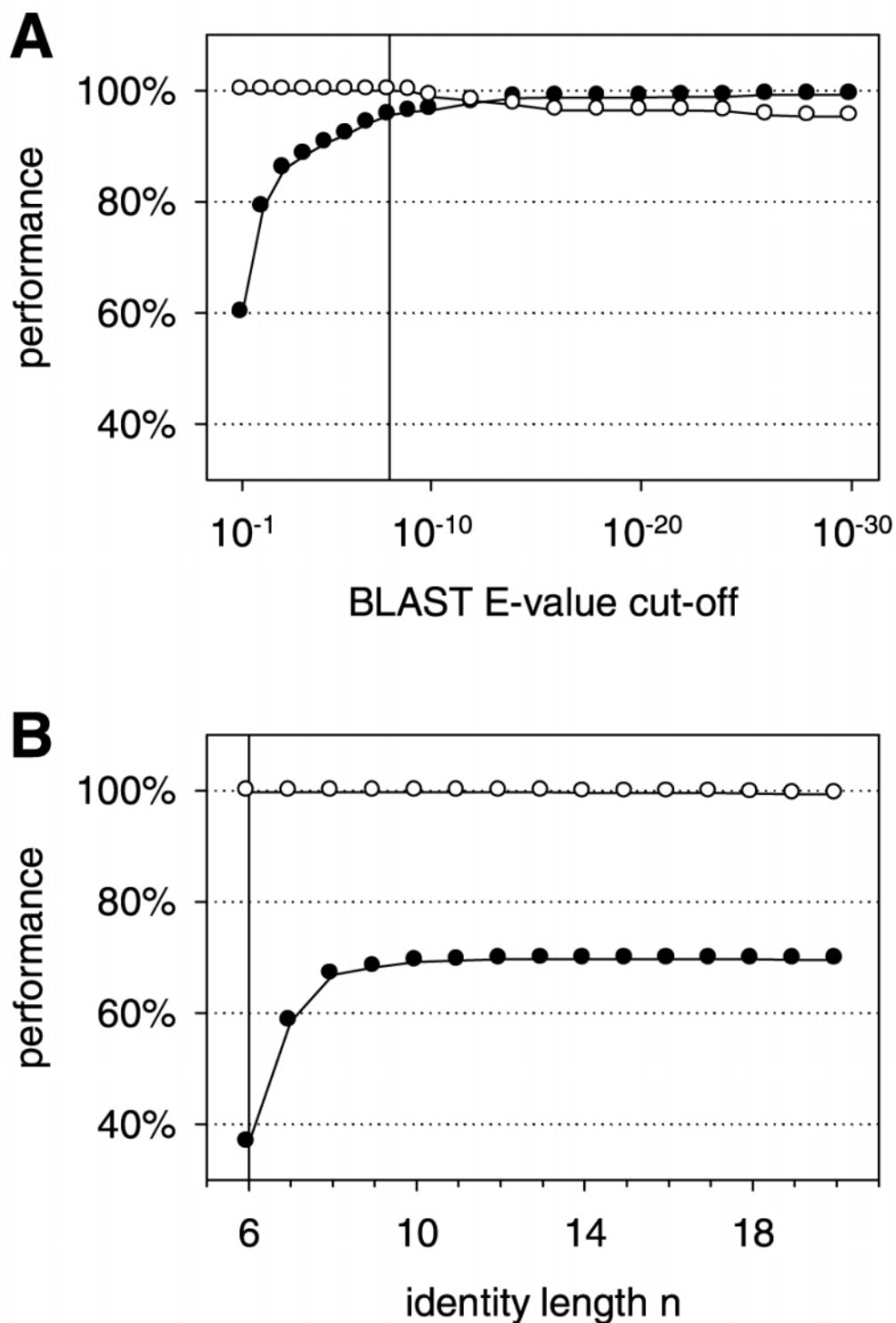


Figure 1. Performance of motif based allergenicity prediction and of prediction according to FAO/WHO guidelines. A test database containing 2,976 protein sequences and 25 % true allergens has been submitted to allergenicity prediction by motif based prediction strategy (**A**) and according to FAO/WHO guidelines (**B**), using a lower length limit of 25 residues for both methods. Prediction stringency has been varied by BLAST E-value cut-off (used in second analysis step of motif based prediction) and identity length for FAO/WHO prediction method. Prediction performance is shown as % precision (●) and % recall (○). The vertical lines indicate BLAST E-value cut-off 10^{-8} and identity length $n=6$.