



Predicting allergenic proteins using wavelet transform

Kuo-Bin Li, * Praveen Issac and Arun Krishnan

Bioinformatics Institute, 30 Biopolis Street, Singapore 138671, Singapore

Received on December 5, 2003; revised on April 7, 2004; accepted on April 13, 2004

Advance Access publication April 29, 2004

ABSTRACT

Motivation: With many transgenic proteins introduced today, the ability to predict their potential allergenicity has become an important issue. Previous studies were based on either sequence similarity or the protein motifs identified from known allergen databases. The similarity-based approaches, although being able to produce high recalls, usually have low prediction precisions. Previous motif-based approaches have been shown to be able to improve the precisions on cross-validation experiments. In this study, a system that combines the advantages of similarity-based and motif-based prediction is described.

Results: The new prediction system uses a clustering algorithm that groups the known allergenic proteins into clusters. Proteins within each cluster are assumed to carry one or more common motifs. After a multiple sequence alignment, proteins in each cluster go through a wavelet analysis program whereby conserved motifs will be identified. A hidden Markov model (HMM) profile will then be prepared for each identified motif. The allergens that do not appear to carry detectable allergen motifs will be saved in a small database. The allergenicity of an unknown protein may be predicted by comparing it against the HMM profiles, and, if no matching profiles are found, against the small allergen database by BLASTP. Over 70% of recall and over 90% of precision were observed using cross-validation experiments. Using the entire Swiss-Prot as the query, we predicted about 2000 potential allergens.

Availability: The software is available upon request from the authors.

Contact: kuobin@bii.a-star.edu.sg

INTRODUCTION

Allergens are proteins that induce allergic responses. More specifically, they elicit immunoglobulin E (IgE) antibodies and cause the symptoms of allergy, which has been a major health problem in developed countries (Ono, 2000).

With many transgenic proteins introduced into the food chain, the need to predict their potential allergenicity has become a crucial issue. Bioinformatics, more specifically,

sequence analysis methods have an important role in the identification of allergenicity (Hileman *et al.*, 2002; Gendel, 2002). To identify allergenicity of a novel protein, World Health Organization (WHO) and the Food and Agriculture Organization (FAO) provides a guideline (Bindsløv-Jensen *et al.*, 2003). In addition to biological tests, this guideline says that a query protein is potentially allergenic if it either has an identity of at least six contiguous amino acids or a minimum of 35% sequence similarity over a window of 80 amino acids when compared with known allergens (Aalberse, 2000). The six amino-acid identity rule is not practical because it produces a large number of false positive hits. The criterion of having a minimum of 35% sequence similarity is on the other hand too stringent to find most true allergens. In any case, it has been shown (Stadler and Stadler, 2003) that the precision is low for methods solely relying on sequence similarities.

Analyses using protein motifs are an emerging method for the prediction of potential allergenicity (Stadler and Stadler, 2003). The methods involve the following steps. First, a reference allergen database is to be created based on known allergens. Then a motif identification tool, such as MEME (Bailey and Elkan, 1994), is used to identify candidate motifs among all proteins in the allergen database. The candidate motifs are then used to create a protein profile. Finally, a query protein is predicted to be allergenic if its sequence matched an allergen motif with a high score.

In this paper, we describe an allergenicity prediction system that appears to be capable of producing high recall and precision. The system is derived from the above-mentioned motif approach but is based on a novel motif detecting tool using wavelet analysis. We have shown that by converting amino acid sequences into numerical ones and performing wavelet analysis on the sequences, conserved motifs can be seen on different scales of the transformed signals (Krishnan *et al.*, 2004). There are two important prerequisites to using this method. First, the input of the motif detection algorithm should be aligned protein sequences. Second, the wavelet method was designed to identify motifs that are common among all sequences. To fulfill the two prerequisites, we must first group the proteins in the allergen database into clusters so that proteins within any cluster share one or more common allergenicity-related motifs. The proteins within each cluster

*To whom correspondence should be addressed.

are aligned using any standard multiple alignment tool and the results were sent to the wavelet analysis program. Hidden Markov model (HMM) profiles were built upon these detected conserved motifs. The proteins that do not share a common motif with others were saved into a small database. This database is assumed to represent the unique allergens. A query protein has been predicted to be allergenic, if it has a high HMM matching score with any of the previously built profiles, or a high pairwise similarity score with any of the sequences in the unique allergen database.

Performance was measured as precision and recall using 10 cross-validation experiments. Finally, the allergenicity of all the proteins in Swiss-Prot was predicted.

SYSTEMS AND METHODS

Selection of known allergens

The list of known allergens was constructed by searching four allergen-specific data sources. It includes 238 allergens from IUIS (www.allergen.org/List.htm), 270 from Swiss-Prot's Allergen Index (www.expasy.org/cgi-bin/lists?allergen.txt), 1171 from BIFS (www.iit.edu/~sgendel/fa.htm, including food allergen, non-food allergen and wheat gluten sequences) and 752 from FARRP (www.allergenonline.com). Sequences redundant in accession number and in 95% sequence identity levels were removed. The 95% identity threshold was chosen for the homology reduction to obtain a smaller set of allergens in which none of the sequences displace significant similarity. Since the smallest known allergen protein has 26 amino acids (bee venom melittin, Swiss-Prot P01501) (King and Spangfort, 2000), a lower bound on length of 25 amino acids was adopted. This resulted in a reference database of 664 allergen proteins. Their accession numbers can be found in the supplementary file.

Prediction system

The flowchart in Figure 1 describes the method that we had used for allergenicity prediction. The distances between each pair of allergen protein sequences were calculated using ClustalW's global alignment functionality. These distances were then extracted and used to cluster the sequences into different groups so that sequences in a cluster have the same conserved regions. This was done through the partitioning around medoids method (Kaufman and Rousseeuw, 1990) using the R package (Ihaka and Gentleman, 1996). The reference allergen sequences were thus divided into many smaller clusters. The sequences in each cluster were then multiple-aligned using either ClustalW (Thompson *et al.*, 1994) or T-Coffee (Notredame *et al.*, 2000). This step was needed since the motif prediction in the next step required the input sequences to be aligned. Each of the cluster of sequences was then analyzed to identify allergen motifs using the wavelet analysis method and all the allergen motifs were thus extracted. We had found that there was ~20% of the sequences

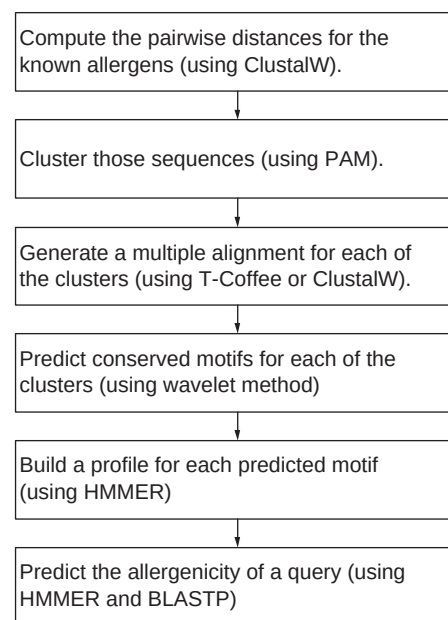


Fig. 1. Flowchart of the entire prediction system.

for which no conserved motifs were detected. To reduce the number of such unique sequences, we implemented a two-step clustering procedure. In the first step, the initial unique sequences are matched against the motifs already obtained. The remaining unique sequences go through a re-clustering procedure to obtain all new motifs. The new motifs were added to the allergen motif list. Following this, an HMM profile for each motif was created using the HMMER package (Eddy, 1998). The allergen protein sequences, belonging to the clusters where no motifs are detected, were grouped together into another database. Any protein sequence can now be tested for allergenicity by scoring it against the database of allergen motif profiles first and then performing a similarity search using BLASTP against all the sequences in the second database. If either of these tests resulted in scores with *E*-values less than a threshold, the query sequence is predicted to be allergenic. Each of the main steps above will be elaborated in the sections below.

Clustering approach

To cluster the protein sequences, the distances between each pair of sequences are required. ClustalW was used to perform multiple alignment of the sequences and the global pairwise distances were extracted. Using these distances, we used R package to cluster the allergen sequences by using the partitioning around medoids method. In this method, at first *K* objects are selected as initial medoids. Then all possible swaps were tried and the *K* objects, which minimized the average dissimilarity between the members of each cluster with their representative object, were selected as the final medoids. All the allergen sequences were thus clustered into *K* clusters.

Wavelet analysis

Wavelet transform is a signal processing method efficient for multi-resolution analysis and local feature extraction of non-stationary signals. It involves decomposing a given signal into its scale and space components. Information can be obtained about both the amplitude of any ‘periodic’ signals as well as when/where it occurred in time/space. Wavelet analysis is thus localized both in space and scale, unlike the Fourier transform in which time/scale information is lost. Although the short-time Fourier transform (STFT) do give a time/scale-frequency, two-dimensional decomposition of the data, its use is limited by the fact that it uses fixed sized windows. In contrast, wavelet transforms use smaller windows for high frequency data and longer windows for low frequency data.

The detailed wavelet approach for the detection of protein motifs is shown elsewhere (Krishnan *et al.*, 2004).

Building HMM profiles

After the wavelet analysis of each cluster, all the motifs that were identified were extracted. HMMER package was used to build profile HMMs of each motif. This was done using the hmmbuild program of the HMMER package.

Implementation

The implementation of all the experiments were carried out using C, Perl and bash scripts. Perl scripts were used to format outputs generated by ClustalW to extract the global pairwise distances and then run them through the R package to cluster the sequences. The wavelet analysis of the protein sequences to identify motifs was implemented using C. Perl scripts were also written to format and extract results after querying the profile database. The complete workflow was implemented as a bash script.

The software is available upon request from the authors.

Cross-validation experiments

A 10-fold cross-validation experiment (Cover, 1969) was carried out to test the performance of our predictions. The allergen database is randomly divided into 10 subsets. Each time one of the 10 subsets is used as the query set and the other nine subsets are put together to form a training set. Every allergen sequence gets to be in a query set exactly once, and gets to be in a training set nine times. The performance was measured as recall and precision. The recall measures the algorithm’s ability to pick out true allergens, while the precision is a measurement of the ability to separate true allergens from the false ones.

$$\text{recall} = \frac{t^+}{t^+ + f^-}, \quad \text{precision} = \frac{t^+}{t^+ + f^+},$$

where t^+ is a true positive, i.e. a true allergen with a score above the prediction threshold; f^+ is a false positive, a false allergen (a randomly shuffled allergen sequence) with a score

above the prediction threshold and f^- is a false negative, a true allergen with a score below the prediction threshold.

RESULTS

The performance of our allergen prediction system is assessed by a 10-fold cross-validation experiment. Here, the 664 reference allergens were randomly divided into a query set containing 66 sequences and a training set containing 598 allergens. The clustering and the wavelet method were applied to the training set to produce conserved allergen motifs. HMM profiles were built based on those predicted allergen motifs. A sequence in the query set is predicted as a potential allergen if it has a high scoring hit with any allergen motifs using hmmpfam. In addition, when the query sequence does not have an hmmpfam hit, the prediction is based on a subsequent BLASTP search against those reference allergens that do not produce any allergen motifs in the previous wavelet analysis. The whole process was repeated 10 times. Each time, a different partitioning was adopted. Table 1 shows the recalls and the precisions of the cross-validation experiments. If the allergenicity of a query protein was predicted using only the hmmpfam program, our system reached an average precision of 99% and a recall of 64%, meaning that none of the shuffled query protein (considered as a false allergen) was declared as a positive and only 64% of the true allergens were predicted as positives. The performance was improved when the BLASTP similarity search was added. Here, a precision of 99% and a recall of 71% were reached. It is worth noting that the recalls increase while performing additional BLASTP searches, the precisions remain roughly the same. Hence calculating of precision does not give us any new insights in view of the lack of knowledge of the actual number of allergens in Swiss-Prot.

To compare the performance with different hmmpfam *E*-value cut offs, we did the cross-validation experiments using three cut offs, 0.1, 0.01 and 0.001. As can be seen in Table 2, a stringent *E*-value threshold improves the precision but decreases the recall. In Table 1, we chose the cut off of 0.001 that produced the highest precision.

Finally, we tested the performance on the entire Swiss-Prot database. The allergenicity was predicted for each of the 135 850 Swiss-Prot proteins (as of 11 November 2003). All 664 reference allergens were used to prepare the allergen motifs. The 62 predicted allergen motifs were listed in Table 3.

The results are summarized in Table 4. A total of 2042 potential allergens are predicted using the motif-based prediction. The list of the 2042 allergens can be found as a supplementary file. Adding BLASTP searches, the number goes up to 4768. To estimate the recall of the prediction, first we noticed that 341 of the 664 reference allergens can be found in Swiss-Prot. Among them, the motif-only method reports 295 while the motif+BLASTP method reports 319 potential allergens. This translates to recalls of 87 and 94% for respective methods. The estimation of the precision is more difficult.

Table 1. Results for the 10-fold cross-validation experiments

	Number of motifs	Number of unique sequences	% Recall Motif only	Motif+BLASTP	% Precision Motif only	Motif+BLASTP
Test1	66	26	66.67	72.73	100.00	100.00
Test2	65	81	59.09	72.73	100.00	100.00
Test3	65	47	69.70	71.21	100.00	100.00
Test4	72	34	66.67	69.70	100.00	100.00
Test5	68	38	66.67	68.18	100.00	100.00
Test6	54	70	59.09	72.73	100.00	100.00
Test7	58	55	60.61	69.70	100.00	100.00
Test8	62	51	66.67	72.73	100.00	100.00
Test9	67	54	56.06	63.64	97.37	97.67
Test10	61	37	65.15	72.73	100.00	100.00
Average	64	50	63.64	70.61	99.74	99.77

The reference allergen sequences are randomly divided into 10 subsets. For each test, the allergenicity of the proteins in a query subset was predicted using the sequences in the remaining nine subsets. Then the sequences in the query subset were randomly shuffled. Any positive prediction from this shuffled query subset was considered as a false positive. Number of motifs is the number of allergen motifs detected by the wavelet methods. Number of unique sequences represents the number of sequences that do not contribute to any predicted allergen motif. For all tests, the number of clusters was set to 50, while the minimal length of a valid allergen motif was set to 30. For the 'motif+BLASTP' experiments, the recalls and precisions were computed using two prediction stages, the profile searching using hmmpfam and the similarity searching using BLASTP. For the 'motif only' experiments, only the first stage was involved. Both the hmmpfam and BLASTP *E*-value cut offs were set to 0.001.

Table 2. Comparison of the allergen predictions using different hmmpfam *E*-value cut offs

	% Recall			% Precision		
	$E < 10^{-3}$	$E < 10^{-2}$	$E < 10^{-1}$	$E < 10^{-3}$	$E < 10^{-2}$	$E < 10^{-1}$
Test1	66.67	66.67	75.76	100.00	97.78	90.91
Test2	59.09	66.67	77.27	100.00	93.62	91.07
Test3	69.70	71.21	75.76	100.00	95.92	94.34
Test4	66.67	68.18	69.70	100.00	97.83	92.00
Test5	66.67	69.70	75.76	100.00	93.88	90.91
Test6	59.09	68.18	74.24	100.00	95.74	87.50
Test7	60.61	65.15	71.21	100.00	97.73	88.68
Test8	66.67	72.73	74.24	100.00	96.00	89.09
Test9	56.06	57.58	60.61	97.37	90.48	86.96
Test10	65.15	68.18	75.76	100.00	100.00	96.15
Average	63.67	67.43	73.03	99.74	95.90	90.76

The 10-fold cross-validation experiments are identical to that of Table 1. For all tests, the number of clusters was set to 50. The minimal length of a valid allergen motif was set to 30.

Although we found that 849 proteins in Swiss-Prot carry the keyword allergen in the annotation, the actual number of true allergens in Swiss-Prot might be higher than that since many proteins have unknown allergenicity.

DISCUSSION

The first task of our allergenicity prediction system is to group the reference allergens into clusters in such a way that sequences within a cluster are to share one or more common motifs. This clustering procedure is a prerequisite for our motif detection approach, because the wavelet method attempt to find conserved regions among all input sequences. The clustering algorithm used in our system is the partitioning

around medoids method (PAM) (Kaufman and Rousseeuw, 1990). Being a partitioning method, it constructs K clusters in the sense that sequences of the same cluster should be related to each other. ClustalW was used to provide the global similarity information as the criterion of the partitioning. It is important to note that K is a user-supplied parameter. Large K indicates small cluster size, i.e. sequences having common allergen motifs might not be placed in a cluster by the PAM algorithm. This reduces the number of effective allergen motifs and thus produces lower recall. On the other hand, small K (large cluster size) will cause some of the irrelevant proteins being placed into one cluster, which also reduces the number of effective allergen motifs. Figure 2A shows the dependency of the recalls versus the number of clusters, K . A compromised value of 50 was chosen for data shown in Table 1. The precision does not vary considerably, this is because true and false positives will both be affected by the reduced number of correctly detected allergen motifs.

As described in the Wavelet analysis section, to determine conserved motifs, the program looks for regions with a minimum of l continuous residues whose signal satisfies some conditions. In all the above tables, the parameter l was set to 30. Figure 2B shows the dependency of the prediction performance on the parameter l . Setting a smaller l would make the wavelet method generate more conserved motifs and thus more true positive predictions (higher recall), at a cost of a slightly lower precision. However, a very small l may produce fragmented allergen motifs that could hurt the recall.

The wavelet and motif-based approach assumes that the allergenicity can be better predicted by using allergen-derived motifs than using predefined protein signatures

Table 3. Allergen motifs

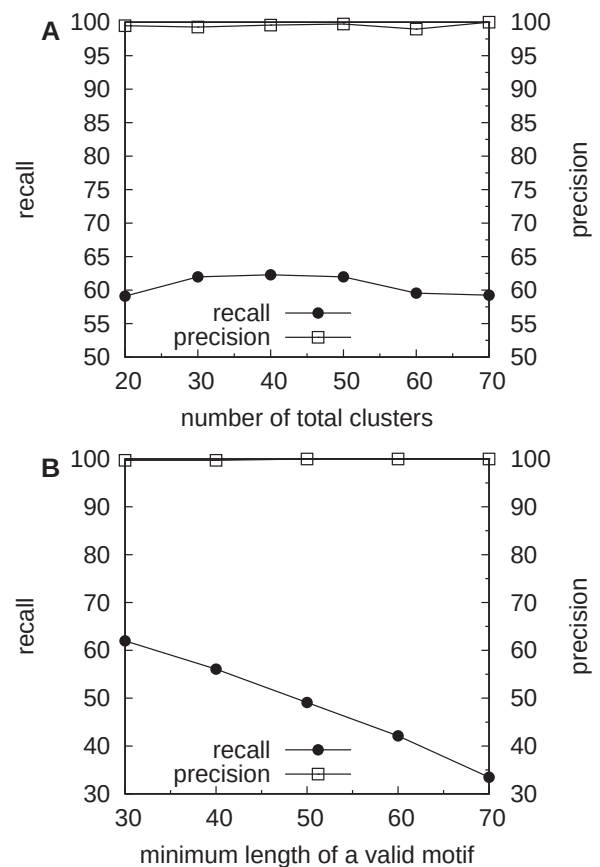
Motif identifiers	Main protein families
motif-1	Profilin
motif-2	Glutenin low molecular weight
motif-3	Agglutinin
motif-4	Glutenin high molecular weight
motif-5	Glutenin high molecular weight
motif-6	Beta-expansin
motif-7	Major mite fecal allergen
motif-8	Profilin
motif-9	Beta-fructofuranosidas
motif-10	Seed allergenic protein
motif-11	Profilin
motif-12	Aspartic protease inhibitor
motif-13	Tropomyosin
motif-14	Tropomyosin; chlorophyll a-b binding proteins
motif-15	Tropomyosin; heat shock protein
motif-16	Gamma-gliadin
motif-17	Profilin
motif-18	Cysteine protease inhibitor
motif-19	Cysteine protease inhibitor
motif-20	Alpha/beta-gliadin
motif-21	Polygalacturonase
motif-22	Gamma-gliadin precursor
motif-23	Polygalacturonase
motif-24	Polcalcin
motif-25	Patatin precursor
motif-26	Patatin precursor
motif-27	Aldehyde dehydrogenase
motif-28	Hyaluronoglucosaminidase precursor
motif-29	Alpha-amylase inhibitor
motif-30	13S globulin seed storage protein 3 precursor
motif-31	13S globulin seed storage protein 3 precursor
motif-32	Lymphocyte antigen 75 precursor
motif-33	Glutenin, high molecular weight
motif-34	Paramyosin
motif-35	Glutenin, high molecular weight
motif-36	Glutenin, high molecular weight
motif-37	Enolase
motif-38	Glycinin
motif-39	Glycinin
motif-40	Gamma-gliadin B-I precursor
motif-41	Calreticulin precursor
motif-42	Beta-conglycinin
motif-43	Glycinin precursor
motif-44	Pollen allergen Amb a 5
motif-45	Parvalbumin
motif-46	Globin
motif-47	Superoxide dismutase
motif-48	13S globulin seed storage protein 3 precursor
motif-49	Profilin
motif-50	Profilin
motif-51	Allergen Pha a 1 precursor
motif-52	Beta-expansin 1 precursor
motif-53	Allergen Hol 1 1 precursor
motif-54	Patatin T5 precursor
motif-55	Keratin
motif-56	Phospholipase A1
motif-57	Lipid-transfer protein
motif-58	Melittin precursor
motif-59	Pathogenesis-related protein
motif-60	Glutenin, low molecular weight
motif-61	Allergen Amb a 1.1 precursor
motif-62	Phospholipase A1

The protein families were identified by using *hmmsearch* to search the Swiss-Prot with a profile HMM generated from the corresponding allergen motif.

Table 4. The motif-based allergenicity predictions for the 135 850 Swiss-Prot proteins

Prediction methods	Predicted allergen	% Ratio in Swiss-Prot	Predicted known allergen	% Recall
motif+BLASTP	4768	3.5	319	93.6
motif only	2042	1.5	295	86.5

Known allergens are defined as the sequences that can be found both in our reference allergen database (664 sequences) and in Swiss-Prot, using 100% sequence identity as the criterion. We found 341 known allergens. The recalls are the ratios between the predicted known allergens and the 341 total known allergens. The 'motif+BLASTP' method involves subsequent BLASTP searches, *E*-value of 10^{-3} was used.

**Fig. 2.** The prediction performance versus two parameters: (A) the number of total clusters and (B) the minimum length of a valid motif. The data were collected using the 10-fold cross-validation experiments without doing BLASTP search.

as those in PROSITE (Falquet *et al.*, 2002). The results in Table 5 demonstrate that some of the allergen-derived motifs, in the example, motif-14 and motif-55, are indeed potential allergenicity determinants rather than protein family signatures. In particular, motif-14 correctly covers two IgE binding epitopes (Ivancic *et al.*, 2002)

Table 5. Allergenic and non-allergenic tropomyosin

Non-allergenic tropomyosin	Matched motif	Allergenic tropomyosin	Matched motif
gi:20522235	motif-15, motif-25, motif-26 motif-30, motif-31	gi:14423954	motif-13, motif-14, motif-15, motif-16, motif-26, motif-30, motif-31, motif-55
gi:45382323	motif-15, motif-26, motif-30 motif-31	gi:42559556	motif-13, motif-14, motif-15, motif-25, motif-26, motif-29, motif-30, motif-31, motif-55,
gi:27465053	motif-15, motif-26, motif-30 motif-31	gi:607633	motif-13, motif-14, motif-15, motif-25, motif-26, motif-30, motif-31, motif-55

The allergenic tropomyosins are from dust mite (gi:14423954), *Chironomus kiiensis* (gi:42559556) and *Metapenaeus ensis* (gi:607633). The non-allergenic tropomyosins are from chicken (gi:20522235), chicken (gi:45382323) and chicken (gi:27465053). In the example, motif-14 and motif-55 are likely to be epitope-related, whereas motif-15, motif-26, motif-30 and motif-31 are likely to be protein family signatures.

VAALNRRIQLLEEDLERSEER and RSLSDDEERMDALENQ. Other motifs in Table 5, on the other hand, are likely to be protein family signatures. Using a shorter length in the wavelet motif detection stage is expected to result in more epitope-related motifs. However, it might also results in more false positive predictions.

An improved prediction performance can generally be obtained by including additional BLASTP similarity searches, as seen in Table 1. The BLASTP searches are performed against the unique sequences, defined as the sequences in the clusters where no conserved motif can be found by our motif detection algorithm. The unique sequences represent those proteins that do not contribute to any allergenicity-related motif. To reduce the number of such unique sequences, we implemented a two-step clustering procedure. In the first step, the initial unique sequences are matched (using hmmpfam) against the motifs already obtained. This is to reduce the impact of incorrect clustering. The remaining unique sequences go through a re-clustering procedure to obtain all new motifs. The new motifs are added to the allergen motif list. Table 1 shows that the average number of unique sequences is 50, i.e. ~8% of the allergens (598 sequences) used in the cross-validation tests.

It is known that the amino acids essential for antibody binding are not necessarily contiguous (Gendel, 2002). Hence a simple similarity search, like the BLASTP we used, might not be efficacious to assess all potential allergenicity. The current recall, 82.47% by using simple BLASTP, might be improved if more sophisticated procedures such

as the multi-step similarity searchings (Gendel, 2002) are adopted.

The FAO approach, an identity of six continuous amino acids with a known allergen, has been reported to be able to produce a high recall of 97% but a low precision of 38% (Stadler and Stadler, 2003). By increasing the length of the continuous amino acids to eight, the recall drops to 92% while the precision goes up to 68%. In addition, the FAO approach predicted that 67% of the Swiss-Prot proteins as potential allergens. This is an indication that the 6mer matching criterion is likely to produce considerable random matches. Comparing our prediction method with the earlier motif-based allergenicity predicting approach (Stadler and Stadler, 2003), although not using the same reference data, both methods produced similar recall and precision. The earlier approach has a predicted 2603 potential allergens for the 101 602 Swiss-Prot proteins. In comparison, our method reports 2042 potential allergens for the 135 850 Swiss-Prot proteins. The earlier approach used MEME as the motif detection tool whereas our method is based on our own wavelet transform. Given a set of sequences, MEME requires users to specify the maximum number of motifs to be identified. As a result, to find all motifs of the reference allergens, an iterative process must be used where the matching motifs are removed from the query set and the remaining sequences go to the next iteration as a new query. Our wavelet approach is conceptually different. The query sequences are partitioned into clusters; then the sequences in each cluster are aligned and wavelet transformed into levels with different resolutions. Conserved motifs are identified on these multi-resolution plots.

To apply our method to the prediction of food allergens, it may be more important to trade-off precision in order to achieve higher recall, as the consequences of missing a true allergen are more severe than false positive predictions. Several options can be done for this purpose. First, the recall can be increased by relaxing the HMM searching threshold. In the cross-validation experiments, the increase in the *E*-value of hmmpfam from 0.001 to 0.1 improves the corresponding recall from 63 to 73%, at the expense of a lower precision from 99 down to 91%. Second, making the minimum length of a valid motif shorter would make the wavelet method generate more predicted motifs, which corresponds to more true positives and false positives. Finally, our motif-based approach may be combined with a homology-based approach. As the results shown in Table 4, an additional BLASTP searching for these query proteins that do not match any predicted allergen motifs would increase the recall from 86 to 93%.

The ability to identify allergic cross-reactivity is important in computational analysis of allergens. Cross-reactivity occurs when the immune system mistakes a similar protein for an allergen. With more annotated allergens added into the protein databases, our method may have a potential of detecting cross-reactivity over distantly related allergens.

ACKNOWLEDGEMENTS

Chua Gek Huey and Wang Yi are thanked for useful discussions. Dr Francis Tang is acknowledged for his assistance on computation tasks. This work was supported by the Bioinformatics Institute, Singapore.

REFERENCES

- Aalberse, R.C. (2000) Structural biology of allergens. *J. Allergy Clin. Immunol.*, **106**, 228–238.
- Bailey, T.L. and Elkan, C.P. (1994) Fitting a mixture model by expectation-maximization to discover motifs in biopolymers. *Proceedings of the 2nd International Conference on Intelligent Systems for Molecular Biology*, Menlo Park, CA. AAAI Press, pp. 28–36.
- Bindslev-Jensen, C., Sten, E., Earl, L.K., Crevel, R.W.R., Bindslev-Jensen, U., Hansen, T.K., Skov, P.S. and Poulsen, L.K. (2003) Assessment of the potential allergenicity of ice structuring protein type III HPLC 12 using the FAO/WHO 2001 decision tree for novel foods. *Food Chem. Toxicol.*, **41**, 81–87.
- Cover, T.M. (1969) Learning in pattern recognition. In Watanabe, S. (ed.), *Methodologies of Pattern Recognition*. Academic Press, New York, pp. 111–132.
- Eddy, S.R. (1998). Profile hidden Markov models. *Bioinformatics*, **14**, 755–763.
- Falquet, L., Pagni, M., Bucher, P., Hulo, N., Sigrist, C.J.A., Hofmann, K., and Bairoch, A. (2002) The PROSITE database, its status in 2002. *Nucleic Acids Res.*, **30**, 235–238.
- Gendel, S.M. (2002) Sequence analysis for assessing potential allergenicity. *Ann N.Y. Acad. Sci., USA*, **964**, 87–98.
- 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.
- Ihaka, R. and Gentleman, R. (1996) R: a language for data analysis and graphics. *J. Comput. Graph. Stat.*, **5**, 299–314.
- Ivanciuc, O., Schein, C.H. and Braun, W. (2002) Data mining of sequences and 3D structures of allergenic proteins. *Bioinformatics*, **18**, 1358–1364.
- Kaufman, L. and Rousseeuw, P.J. (1990). *Finding Groups in Data: An Introduction to Cluster Analysis*. John Wiley & Sons, Brussels, Belgium.
- King, T.P. and Spangfort, M.D. (2000) Structure and biology of stinging insect venom allergens. *Int. Arch. Allergy Immunol.*, **123**, 99–106.
- Krishnan, A., Li, K.-B. and Issac, P. (2004) Rapid detection of conserved regions in protein sequences using wavelets. *In Silico Biol.*, **4**, 0013.
- Notredame, C., Higgins, D.G. and Heringa, J. (2000) T-Coffee: a novel method for fast and accurate multiple sequence alignment. *J. Mol. Biol.*, **302**, 205–217.
- Ono, S.J. (2000) Molecular genetics of allergic diseases. *Annu. Rev. Immunol.*, **18**, 347–366.
- Stadler, M.B. and Stadler, B.M. (2003) Allergenicity prediction by protein sequence. *FASEB J.*, **17**, 1141–1143.
- Thompson, J.D., Higgins, D.G. and Gibson, T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, **22**, 4673–4680.