

Arbeiten aus dem Paul-Ehrlich-Institut
11th International Paul-Ehrlich-Seminar, Bad Homburg 2005
Verlag Chmielorz

Computational Resources: A Regulatory Need, a Tool for Research

Renato Truffer, Michael B. Stadler, Monique Vogel, Adriano Mari,
Beda M. Stadler

Summary

Background: We have recently reported a new method based on protein motifs to predict allergenicity by protein sequence [1]. By using the more complete allergen database Allergome (www.allergome.org) [2], as the basis for the motif identification process, we obtained a new set of allergenic motifs. We studied whether an allergenic motif may be used instead of a full length allergen for predicting allergenicity.

Methods: Sequence accession numbers of allergens were received from A. Mari of the Allergome Project. The sequences were downloaded and integrated in the allergen database, which was used to produce new motifs applicable for allergenicity prediction. The motif containing allergenic tropomyosins of different non vertebrate species was chosen. A consensus sequence of the tropomyosin motif was synthesized and expressed in *Escherichia coli* (*E. coli*). The purified peptide was analyzed for specificity with enzyme linked-immunosorbent assay (ELISA).

Results: Based on the Allergome dataset, we calculated 69 motifs containing 912 sequences. Prediction performance of this new set of allergenic motifs is increased compared to the original 52 motifs. The IgE immunological reactivity of the recombinant tropomyosin motif was comparable to that obtained with a full length recombinant shrimp tropomyosin (rPen a 1) tested in the ImmunoCAP system.

Discussion: The use of the motif-based allergenicity prediction method is more accurate than the guidelines proposed by FAO/WHO [3], however continuous updates of the motif dataset are necessary to guarantee a high prediction performance of this method. The experiments performed with the tropomyosin motif showed that peptides based on the consensus sequence of motifs may represent the allergenic epitope, which can be used in future diagnostic or therapeutic approaches.

Introduction

There is a general need to predict allergenicity of protein sequences. We have recently published a method to analyze protein sequences in order to evaluate their allergenic potential [1]. An allergen database was generated by downloading all sequences of various published allergen lists. Redundancies have been removed, resulting in an allergen database containing 779 sequences. Then the motif discovery tool MEME [4] was used to cluster the proteins of the allergen database, leading to 52 motifs and 135 sequences which did not match any motif. These motifs were used to predict the allergenic potential of a query protein sequence. In comparison with the FAO/WHO guidelines for the evaluation of allergenicity of genetically modified foods this new method has a similar recall (sensitivity). However, the precision (percentage of relevant hits in relation to the total number of hits) of the motif-based prediction (95.5%) is superior to the FAO/WHO method (36.6%) [1].

In this study we repeated this analysis using the Allergome database (www.allergome.org) [2], which is updated continuously. The last motif-based analysis was performed on 1102 allergenic sequences, resulting in 69 motifs, 190 sequences could not be assigned to any motif. As it has been reported that most of the clinical relevant allergens are known [5], it can be assumed that the number of new allergenic sequences will not increase much more and that the total number of motifs may remain below 100.

In order to evaluate the clinical significance of the described allergenic motifs, a 50 amino acid long consensus sequence of the tropomyosin motif was synthesized and expressed in *E. coli*. The purified peptide was tested with ELISA on sera of allergic patients and showed the same reactivity as that obtained with a full length recombinant shrimp tropomyosin (rPen a 1) used in an ImmunoCAP test.

Results and Discussion

The FAO/WHO guidelines for the evaluation of allergenicity of genetically modified foods published in 2001 contain besides biological tests a standard method for sequence comparison: A query protein is potentially allergenic if it either shares more than 35% sequence similarity over a window of 80 amino acids or has an identity of at least six contiguous amino acids with a known allergen. This method applied to an allergen database classifies 98.6% of its sequences as allergenic, but also between 42.9% and 75.9% of the sequences of the Swiss-Prot database, the rice genome and the translated human genome (Table 1). However, according to the

Database	Adb	Swiss-Prot	Rice	trGEN (human)
Description	Allergen database	General protein database	Rice genome (TIGR OsGI)	Human genome (translated)
# Proteins	779	101602	10891	330743
Allergens (predicted)	98.6 %	67.3 %	75.9 %	42.9 %

Table 1. Allergen prediction according to FAO/WHO guidelines

Swiss-Prot allergen index only 0.4% of the proteins are expected to be real allergens. This implies that the guideline described by FAO/WHO is not accurate enough to support the evaluation of allergenicity.

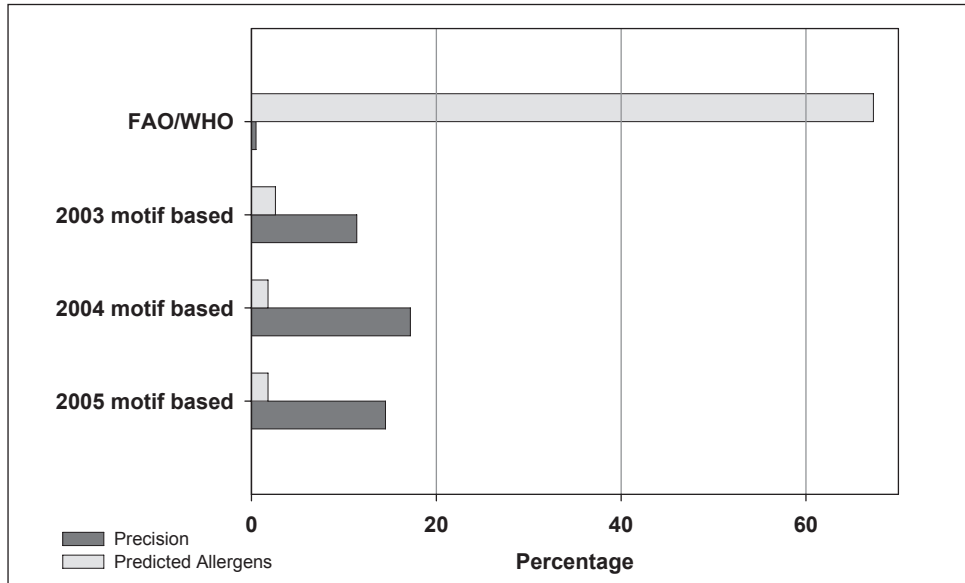
To find a better approach we developed a method to group proteins into motifs. This method based on an automated iterative motif discovery process was first applied on a self-made allergen database [1]. From this database of 779 allergenic sequences we found 52 motifs. Only 135 sequences did not match to any motif. In this study we applied this method using a more complete database provided by A. Mari from www.allergome.org. In the last motif identification process 912 out of 1102 allergenic sequences were clustered into 69 motifs. 190 sequences did not match any motif (Table 2).

	2003	2004	2005
Sequences (origin)	Self made allergen database	Allergome	Allergome
Sequences (#)	779	876	1102
Motifs (#)	52	64	69
Non-matching sequences	135	169	190

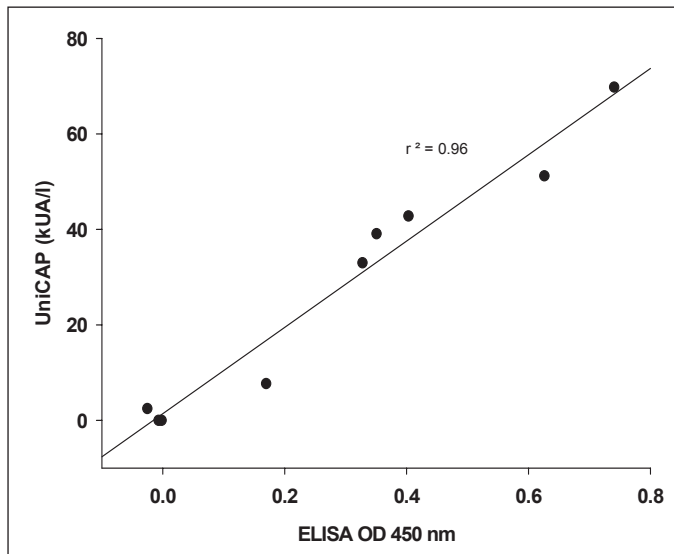
Table 2. Motif identification using different databases

Prediction performance of three different motif datasets (based on different allergen reference databases) was compared to the FAO/WHO method (Figure 1). As reported earlier, the FAO/WHO method classified 67.3% of all proteins in Swiss-Prot as potential allergens. The different motif based predictions reported only two to four percent of the Swiss-Prot proteins as potential allergens, which is more realistic. The precision for all three motif based predictions is remarkably higher than the precision of the FAO/WHO method. The precision values are expressed as the percentage of true allergens in the total predicted allergens. A potential allergen was considered a true allergen, if its sequence was contained in the allergen reference database.

To assess the biological significance of the motifs we chose a motif containing allergenic tropomyosins of different invertebrates. The sequences from this motif were aligned and a consensus sequence of the motif region was determined. This consensus sequence was synthesized as DNA fragments using overlapping oligonucleotides. The DNA fragments were then expressed as recombinant proteins fused to a His₆ tag in *E. coli* (rTropMot). The immunogenicity of the allergenic rTropMot was tested using sera of patients allergic to shellfish. This reactivity was comparable to that obtained with a recombinant full length shrimp tropomyosin (rPen a 1) tested in the ImmunoCAP system (Figure 2). These results support the idea to use motifs instead of whole allergens sequences for future diagnostic procedures.

Fig. 1.: Prediction performance of the motif based method compared to the FAO/WHO method

The Swiss-Prot database (release 40) was analyzed with the FAO/WHO method and three different motif datasets. The percentages of the predicted allergens in Swiss-Prot are represented by light grey bars. The precision values represented by dark grey bars are calculated: $\text{true}^+ / (\text{true}^+ + \text{false}^+)$.

**Fig. 2.:** Correlation of IgE reactivity between recombinant full length tropomyosin (rPen a 1) and recombinant tropomyosin motif (TropMot)

Nine individual shrimp allergic sera (black points) diluted 1:2 were analyzed using ImmunoCap® tests for recombinant full length tropomyosin (rPen a 1). The same individual sera were tested against recombinant tropomyosin motif (rTropmot) in ELISA. Wells were coated with 5 µg/ml rTropmot and incubated with sera diluted 1:2. Bound antibodies were detected using horseradish-peroxidase labeled polyclonal anti-IgE antibodies. The regression line represents the correlation between the two tests.

Conclusion

The analysis of allergen databases using the motif based approach showed a higher degree of prediction performance than the guidelines for the evaluation of the allergenicity of genetically modified foods proposed by the FAO/WHO.

The similar IgE reactivity of the tropomyosin motif compared to the one of the full length tropomyosin suggest that the motifs may be useful to predict allergenicity and may be used in future diagnostic tests.

References

1. Stadler MB, Stadler BM. Allergenicity prediction by protein sequence. *FASEB J* 2003; 17: 1141-1143.
2. Mari A, Riccioli D. The Allergome web site - a database of allergenic molecules. Aim, structure, and data of a web-based resource. *J Allergy Clin Immunol* 2004; 113: S301.
3. FAO/WHO. Evaluation of allergenicity of genetically modified foods. Report of a joint FAO/WHO expert consultation on allergenicity of foods derived from biotechnology. 2001: Food and Agriculture Organization (FAO).
4. Bailey TL, Elkan C. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc Int Conf Intell Syst Mol Biol* 1994; 2: 28-36.
5. Valenta R, Kraft D. From allergen structure to new forms of allergen-specific immunotherapy. *Curr Opin Immunol* 2002; 14: 718-727.

Discussion

Seppälä: You mentioned at the beginning of your talk that this approach would be a good method to predict e.g. the allergenicity of proteins from GMOs. Have you done any work along this line?

Truffer: No, we have not done it yet.

Chapman: Tropomyosins usually have a fairly high degree of sequence homology. Have you applied this motif to group the proteins into allergens that have similar function, but much less sequence homology? A typical example would be the lipocalins which only have 20-30% sequence homology, but a very similar three-dimensional structure. Does it work if you apply your approach to that kind of situation?

Truffer: There are ongoing experiments on that, but we do not have results yet.

Meno: What happens if you have a protein which has structural homologues which are not allergens, like Der p 1 for example which is a cysteine protease? Would that end up with the 135 proteins which are not in a sequence motif?

Truffer: I do not know, actually.

Larsen: Maybe I missed something, if I understood correctly you took out the allergen sequences and then from that extracted motifs which were based on primary sequences. Then you went back to the SwissProt database and identified the allergens?

16

Truffer: Yes, we used all these motifs, we calculated to predict the allergenic proteins in the SwissProt data base and 4% of the proteins in the database were rated as allergens with this method.

Larsen: But are these proteins you find not the same as the ones you originally started with?

Truffer: You find much more proteins because the idea is to find similar proteins, too. The precision of this motif of this search was about 15%. That means that only 15 out of 100 predicted allergens really were allergens, of the other 85 we do not know. We have to do more studies.

Larsen: But if you do not have a sequence or structural aspect to distinguish an allergen, then something else must select which proteins are allergens. Maybe the way they are exposed to humans? Does it make sense to try and predict an allergen based on sequence if we know that sequence is not an aspect in the selection of an allergen?

Truffer: It is a part of the process to identify proteins as allergens. As the first step, you can look at the protein sequence and if you find similarities to an allergen then it might be an allergen. If you do not find any, it is probably not an allergen if the database of allergens you search is big enough.