Pol. J. Food Nutr. Sci., 2012, Vol. 62, No. 3, pp. 151-157
DOI: 10.2478/v10222-011-0036-2
http://journal.pan.olsztyn.pl

Original Paper
Section:

Evaluation of *In Silico* Prediction Possibility of Epitope Sequences Using Experimental Data Concerning Allergenic Food Proteins Summarised in BIOPEP Database

Piotr Minkiewicz, Jerzy Dziuba, Małgorzata Darewicz, Justyna Bucholska, Damir Mogut*

*University of Warmia and Mazury in Olsztyn, Department of Food Biochemistry,
Plac Cieszyński 1, 10–726 Olsztyn-Kortowo, Poland*

Key words: allergy, BIOPEP database, proteins, epitopes, allergen-representative peptides, EVALLER application

The aim of the study was to evaluate the possibility of predicting potential epitope sequences and location in allergenic proteins from food using EVALLER program by comparison with experimental epitopes summarised in the BIOPEP database of allergenic proteins. Sequences of experimental epitopes from food allergens, present in the BIOPEP database of allergenic proteins were used in the study. Sequences of potential epitopes were found using EVALLER program. The Positive Predictive Value (PPV) has been used as a measure of prediction quality. The potential epitopes fully or partially overlapping with the experimental ones were considered as true positive results whereas these unrelated to the experimental ones as false positive results. The PPV for entire dataset containing 310 potential epitopes was 80.6%. The PPV varied significantly among particular allergen families defined according to the AllFam database. Caseins revealed PPV=100% (with one exception), proteins from tropomyosin family and proteins from papain-like cystein protease family – exceeding 50%. The last two families possess also relatively low frequency of epitope occurrence. The predictive potential was poor (less than 50%) for plant allergens from cupin superfamily. Families such as lipocalins from milk and EF-hand family (parvalbumins) revealed high variability within family. The EVALLER program may be used as a tool for the prediction of epitope location although its potential varies considerably among allergen families. High PPV is associated with a high number of known experimental epitopes (such as in caseins) and/or a high degree of sequence conservation within family (caseins, tropomyosins).

INTRODUCTION

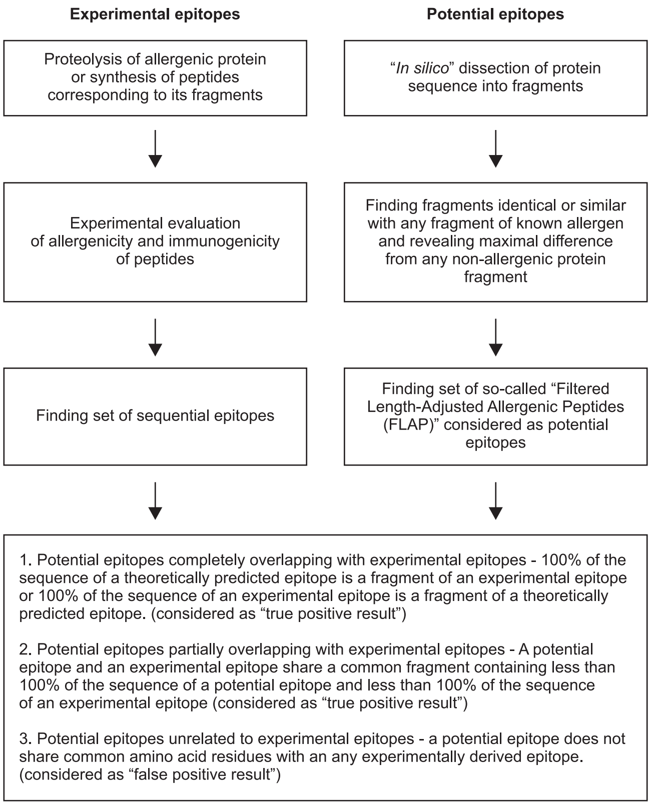
Allergy is one of the greatest challenges for the contemporary food science and medicine [Skripak & Sampson, 2008; Jędrychowski *et al.*, 2008; Cianferoni & Spergel, 2009; Cummings *et al.*, 2010]. Allergy treatment involves the elimination of allergens from the diet or desensitization therapy [Skripak & Sampson, 2008; Kim & Sicherer, 2010; Prescott *et al.*, 2010]. For elimination diets and desensitization treatment to be effective, allergenic food proteins and protein fragments that are epitopes have to be determined. The relevant research is supported by bioinformatics tools [Tong & Ren, 2009; Mari *et al.*, 2009; Salimi *et al.*; 2010; Tomar & De, 2010]. The bioinformatics tools applied in immunology and allergology include databases containing allergenic protein sequences [Gendel, 2009; Mari *et al.*, 2009; Salimi *et al.*, 2010; Tomar & De, 2010; Darewicz *et al.*, 2011]. Those databases are compatible with applications comparing a given protein sequence with the sequences listed in the database. Most of such applications rely on BLAST [Altschul *et al.*, 1997] and FASTA [Pearson *et al.*, 1991; Pearson, 2000] algorithms. The existing databases support the evaluation of a protein's potential allergenicity based on a set of bioinformatics criteria recommended by the World Health Organization (WHO), such as the presence of protein sequence fragments containing a minimum of 6–8 amino acid residues which are identical to the fragments of known allergens or fragments containing a minimum of 80 amino acid residues showing at least 35% similarity with the known allergen sequence [Goodman, 2006]. The search for new solutions is still necessary to further advancement of the existing bioinformatics methods and tools [Gowthaman & Agrewala, 2009].

The list of applications that predict the allergenicity of proteins includes EVALLER program for predicting the allergenicity and cross-reactions of proteins [Martinez-Barrio *et al.*, 2007]. It compares protein sequence fragments with a database of filtered length-adjusted allergen peptides (FLAPs) [Soeria-Atmadja *et al.*, 2006], an improved version of the database of allergen-representative peptides developed by Björklund and coworkers [2005]. The above program searches for fragments in allergenic proteins, and it accounts for differences between those fragments and the sequences of non-allergenic proteins. The overlapping segments, selected based on the above criteria, were combined to form longer fragments – FLAPs. The EVALLER database comprises sequences of 762 allergenic proteins and twice more non-allergenic proteins. EVALLER relies on two criteria for evaluating the degree of sequence matching: the degree of identity expressed by the percentage of identical residues with the same position in a protein chain fragment, which accounts for

* Corresponding author: Tel.: +4889 523 37 15
E-mail: jerzy.dziuba@uwm.edu.pl (Prof. J. Dziuba)

insertion and deletion computed using the FASTA algorithm [Pearson, 1991; 2000], and the Smith-Waterman score [Smith & Waterman, 1981]. The algorithm applied to develop the EVALLER program was compared with other algorithms investigating the allergenic character of the examined proteins [Soeria-Atmadja *et al.*, 2006]. In reference to the official bioinformatics criteria recommended by the WHO [Goodman, 2006], the discussed program produces fewer false-positive results, *i.e.* cases in which a non-allergenic protein is found to be an allergen [Soeria-Atmadja *et al.*, 2006]. According to a suggestion presented by Björklund *et al.* [2005], peptides characteristic for allergens could overlap with experimental epitopes determined by mapping, *i.e.* based on interactions between the peptides corresponding to fragments of protein sequences and the antibodies of persons allergic to the analysed protein [Bohle, 2006; Steckelbroeck *et al.*, 2008]. Comparison of the epitope determination using experimental mapping and prediction using EVALLER is presented in Figure 1.

The aim of the present study was to evaluate the possibility of predicting potential epitope sequences and location in allergenic proteins from food using EVALLER program by comparison with experimental epitopes summarised in the BIOPEP database of allergenic proteins.



Proteolysis of allergenic protein or synthesis of peptides corresponding to its fragments

"*In silico*" dissection of protein sequence into fragments

Experimental evaluation of allergenicity and immunogenicity of peptides

Finding fragments identical or similar with any fragment of known allergen and revealing maximal difference from any non-allergenic protein fragment

Finding set of sequential epitopes

Finding set of so-called "Filtered Length-Adjusted Allergenic Peptides (FLAP)" considered as potential epitopes

1. Potential epitopes completely overlapping with experimental epitopes - 100% of the sequence of a theoretically predicted epitope is a fragment of an experimental epitope or 100% of the sequence of an experimental epitope is a fragment of a theoretically predicted epitope. (considered as "true positive result")

2. Potential epitopes partially overlapping with experimental epitopes - A potential epitope and an experimental epitope share a common fragment containing less than 100% of the sequence of a potential epitope and less than 100% of the sequence of an experimental epitope (considered as "true positive result")

3. Potential epitopes unrelated to experimental epitopes - a potential epitope does not share common amino acid residues with an any experimentally derived epitope. (considered as "false positive result")

FIGURE 1. Scheme of the prediction of potential epitope location using EVALLER program and comparison with experimental epitopes.

METHODS

Information about bioinformatics tools mentioned in this article, including web addresses and references (if available) is summarised in Table 1. All databases and programs were accessed before 30.04.2011.

Sequences of food allergens and experimental epitopes present in the BIOPEP database of allergenic proteins were used in the study.

The BIOPEP database of allergenic proteins and their epitopes contains the following information: allergen name; sequence, sequences of experimental epitopes; sequences of predicted epitopes; reference describing proteins sequence (bookmark "reference"); reference concerning allergenicity, information about sequences of experimental and theoretical epitopes; ID of particular experimental epitopes in the Immune Epitope Database (bookmark "additional information"); information about AllFam allergen family and epitopes occurring in more than one protein in the BIOPEP database (bookmark "homology") and information about annotation in general protein database (mainly UniProt), WHO-IUIS (World Health Organization – International Union of Immunological Societies) and Allergome or other allergen databases (bookmark "database reference"). The BIOPEP database of allergenic protein and their epitopes contains also a search engine that enables finding fragments identical with epitopes in protein sequences (*e.g.* provided by user).

Sequences of potential epitopes were found using EVALLER program with default parameters. Number of sequences displayed in an output covered all sequences revealing 100% identity with fragments of known allergens (as recommended

TABLE 1. Bioinformatics tools mentioned in this publication.

Tool	Website	Reference
Allergome	http://www.allergome.org/	Mari *et al*. [2006, 2009]
AllFam	http://www.meduniwien.ac.at/allergens/allfam/	Radauer *et al.* [2008]
BIOPEP database of allergenic proteins and their epitopes	http://www.uwm.edu.pl/biochemia	Minkiewicz *et al.* [2011]
EVALLER	http://bioinformatics.bmc.uu.se/evaller.html; http://www.slv.se/en-gb/Group1/Food-Safety/ e-Testing-of-protein-allergenicity/e-Test-allergenicity/	Martinez Barrio *et al.* [2007]
Immune Epitope Database (IEDB-AR)	http://www.immuneepitope.org/	Vita *et al.* [2010]
Pfam	http://pfam.sanger.ac.uk/	Sammut *et al.* [2008]
UniProt	http://www.expasy.org	Jain *et al.*, 2009; The UniProt Consortium [2011]
WHO-IUIS allergen database	http://www.allergen.org/	

by Minkiewicz *et al.* [2011]). All such fragments are annotated in the BIOPEP database. Sequences with a lower degree of identity were not taken into account.

The frequency of occurrence of sequential epitopes (A) has been calculated automatically during introduction of protein data into the BIOPEP database according to equation 1 [Dziuba *et al.*, 2003]. Both experimental and potential epitopes are included during calculation.

$$A = n/N \quad (1)$$

where: n – number of epitopes, and N- number of amino acid residues.

Postive predictive value (PPV) has been used as a parameter estimating the quality of prediction of a potential epitope. The PPV may be considered as likelihood that potential epitope will fully or partially overlap with at least one experimental epitope. The PPV value has been calculated using equation 2 [Pulido *et al.*, 2003].

$$PPV = 100tp/(tp + fp) \quad (2)$$

where: tp – true positive results, and fp – false positive values.

Definitions of true and false positive values are presented in Figure 1.

The total PPV value has been calculated for all proteins containing both potential and experimental epitopes. The PPV value for an individual allergen has been calculated only if the allergenic protein contains more than one potential epitope.

RESULTS AND DISCUSSION

Sixty proteins out of the 135 ones present in the BIOPEP database contain both potential and theoretical epitopes, and 43 of them contain at least two potential epitopes determined by EVALLER program. The data cover 310 allergen-representative FLAPs, of which 187 (60.3%) completely overlapped with the experimental epitopes, 63 (20.3%) partially overlapped with the experimental epitopes, and 60 (19.4%) were unrelated to the experimental epitopes. In line with the adopted sequence coverage criterion, a single FLAP has to be found in a single experimental epitope or a single experimental epitope has to be determined in a single FLAP to be considered as fully overlapping. Such a rule will be maintained in the further discussion. The Positive Predictive Value calculated for the entire dataset (total PPV) is 80.6%.

A group of experimental epitopes forming a continuous protein sequence fragment could overlap with a theoretically-predicted allergenic peptide or a group of theoretically-predicted peptides (FLAPs) forming a continuous protein sequence fragment could overlap with an experimental epitope. An example of the above is an experimental epitope sequence containing residues 131–151 in the precursor of allergen Ara h 3.0101 (BIOPEP ID 47) which overlaps with a group of potential epitopes containing residues 108–135; 124–145; 126–147; 139–163; 144–165 and 149–184, respectively. None of the above FLAPs contains or is a part of an experimental epitope. Based on the above, the discussed FLAPs have been classified as partially overlapping with experimental epitope. All six FLAPs create a continuous fragment of a protein sequence containing 100% experimental epitope amino acid residues. If similar epitope groups were to be classified as completely overlapping, the number of completely overlapping sub-sets would increase to 192 (63.4% of all theoretically-predicted allergenic peptides), and the number of FLAPS that partially overlap with experimental epitopes would be reduced to 51 (16.8% of all theoretically-predicted allergenic peptides). The percentage of unrelated potential epitopes would be thus 19.8%.

Proteins containing at least 2 potential epitopes have been divided into subgroups according to A and PPV value as shown in Table 2. For 27 out of 43 proteins mentioned in this table PPV=100%. It means that all potential epitopes found using EVALLER program in the sequences of proteins mentioned above, fully or partially cover known experimental epitopes. Proteins with A value lower than 0.2 are characterised by PPV from 0 to 100%. The PPV value of proteins possessing A values between 0.2 and 0.6 is between 50 and 100%. All proteins with the frequency of occurrence of sequential epitopes larger than 0.6 are characterised by a positive predictive value of 100%. It means that such proteins do not contain potential

TABLE 2. Classification of allergenic proteins from the BIOPEP database according to PPV concerning overlapping of potential and experimental epitopes and A. A and PPV are defined according to equations 1 and 2 respectively (see Methods section). Allergen names are given according to WHO-IUIS and Allergome database. ID number according to BIOPEP database of allergenic proteins are given in parentheses.

PPV (%)	A < 0.2	0.2 < A < 0.6	A > 0.6
100	Bos d 6 (15); Cha f 1 (62); Cor a 1.0404 (70); Cra g 1 (3); Equ as ALA (116); Gad c 1 (17); Gal d 2 (4); Hom a 1.0102 (63); Ses i 2.0101 (52)	Bub b casein alpha s1 (109); Bub b casein alpha s2 (79); Bub b casein beta (110); Gal d 1.0101 (5); Met e 1 (73); Ovi a casein alpha s1 (103)	Bos d 5 (14); Bos d 8 alpha s1 (9); Bos d 8 alpha s2 (10); Bos d 8 beta (11); Bos d 8 kappa (12); Bos g casein kappa (111); Bos i casein beta (113); Bos i casein kappa (114); Bub b BLG (108); Cap h BLG (100); Ovi a BLG (77); Ran t BLG (128)
50 -99.9	Ara h 3.0101 (47); Ara h arachin 6 (51); Gly m Bd 30 K1 (42); Gly m Bd 30 K2 (71); Hom a 1.0101 (72)	Cap h casein alpha s1 (78); Pan s 1 (64)	-
0.1 -49.9	Ara h 4 (98); Ara h 4.0101 (48); Equ c BLG (122); Fag e 1 (40); Ran e 2.0101 (94); Ses i 3.0101 (53)	-	-
0	Equ as BLG (117); Gad m 1.0201 (92); Gad m 1.0201 (16)	-	-

TABLE 3. Ranges of positive predictive value concerning overlapping of potential and experimental epitopes (PPV) and and frequency of sequential epitope occurrence (A) for allergenic proteins of various origin. A and PPV are defined according to equations 1 and 2 respectively (see Methods section).

PPV (%)	A < 0.2	0.2 < A < 0.6	A > 0.6

100	milk (2 proteins);
	crustacea and molllusca (3 proteins);
	plants (2 proteins);
	fishes and amphibians (1 protein);
	eggs (1 protein)
		milk (4 proteins);
		eggs (1 protein); crustacea and molllusca (1 protein)
			milk (12 proteins)

50 -99.9	plants (4 proteins); crustacea and molllusca (3 proteins)	milk (1 protein); crustacea and molllusca (1 protein)	-

0.1 -49.9	plants (4 proteins);
	milk (1 protein);
	fishes and amphibians (1 protein)
		-	-

0	milk (1 protein); fishes and amphibians (2 proteins)	-	-

TABLE 4. Ranges of positive predictive value concerning overlapping of potential and experimental epitopes (PPV) and frequency of sequential epitope occurrence (A) for allergenic proteins belonging to various homology-based families. ID numbers and names of particular families are taken from AllFam database. A and PPV are defined according to equations 1 and 2 respectively (see Methods section).

PPV (%)	A < 0.2	0.2 < A < 0.6	A > 0.6

100	AF007 EF hand - parvalbumins (1 protein);
	AF016 C-type lysozyme/alpha-lactalbumin family (1 protein);
	AF018 Serpin serine protease inhibitor (1 protein);
	AF050 Prolamin superfamily (1 protein);
	AF054 Tropomyosins (3 proteins);
	AF056 Serum albumins (1 protein)
	AF069 Bet v 1-related protein (1 protein)
		AF013 Kazal-type serine protease inhibitor (1 protein)
		AF054 Tropomyosins (1 protein);
		AF065 Alpha/beta casein (4 proteins)
			AF015 Lipocalins (5 protein);
			AF065 Alpha/beta casein (5 proteins);
			AF085 Kappa-casein (3 proteins)

50 -99.9	AF030 Papain-like cysteine protease (2 proteins) AF045 Cupin superfamily (2 proteins); AF054 Tropomyosins (1 protein)	AF054 Tropomyosins (1 protein); AF065 Alpha/beta casein (1 protein)	-

0.1 -49.9	AF007 EF hand-parvalbumins (1 protein);
	AF015 Lipocalins (1 protein);
	AF045 Cupin superfamily (4 proteins);
	AF054 Tropomyosins (1 protein)
		-	-

0	AF007 EF hand-parvalbumins (2 proteins); AF015 Lipocalins (1 protein)	-	-

epitopes unrelated to the experimental ones. Proteins characterised by high A values contain a high number of epitopes found using experimental mapping strategy. Thus, it is easy to find experimental epitopes overlapping with the potential ones. Low PPV (below 50%) values correspond with low A values. This fact may be explained by the incomplete knowledge about epitopes. In addition to sequential epitopes, allergenic proteins also contain conformational epitopes [Pomés, 2010; Ponomarenko *et al.*, 2011]. They are not mentioned in the BIOPEP database, but may also overlap with FLAPs (potential epitopes) pointed out by EVALLER program.

Ranges of A and PPV values of allergenic proteins of various origin are presented in Table 3. The discussed group includes the highest-risk allergens [Jędrychowski *et al.*, 2008]. Milk proteins form a dominant group among the allergens characterised by high values of A indicating a high number of known experimental epitopes. They were subjected to extensive studies aimed at epitope determination as compared with allergenic proteins from other sources [Vaughan *et al.*, 2010]. Moreover sequences of milk proteins are highly conserved and contain common experimental epitopes as pointed out by Minkiewicz *et al.* [2011]. The experimental epitopes were attributed to bovine milk proteins [Monaci *et al.*, 2006; Vaughan *et al.*, 2010; IEDB-AR database]. The current list of bovine milk epitopes has been markedly enriched as compared with that presented in the above-mentioned review. On the other hand, calculations restricted to the epitopes mentioned in the review of Monaci and coworkers [2006] give overall results (PPV calculated for all potential epitopes) almost the same as these presented above. Another example are tropomyosins of crustaceans and other invertebrates. The IEDB-AR database lists the epitopes of shrimp (*Farfantepenaeus aztecus*) tropomyosin (Pen a 1.0102; BIOPEP ID 76). The above protein was subjected to experimental epitope mapping [Shanti *et al.*, 1993; Reese *et al.*, 1999; 2001; 2005; Ayuso *et al.*, 2002]. All invertebrate tropomyosins listed in the BIOPEP database (ID: 3; 24; 62; 63; 64; 65; 72; 73, 74, 93 and 99) contain epitopes shared with allergen Pen a 1.0102. Epitopes discovered first in molluscan tropomyosins [Ishikawa *et al.*, 1998a;b; 2001], not mentioned in the IEDB-AR database, are also present in this allergen as well as in its homologs.

Ranges of A and PPV values of allergenic proteins belonging to different families are presented in Table 4. Classification

TABLE 5. Allergens containing potential epitopes predicted using EVALLER program, but not containing known, experimentally found sequential epitopes (till the end of April 2011).

Allergen[1]	ID[2]	Number of epitopes	Location of epitopes	Origin of protein	Family name and ID according to AllFam database
Api g 1.0101	57	2	[1–41],[79–135]	plant	Bet v 1-related proteins family (AF069)
Api g 1.0201	58	1	[38–135]	plant	Bet v 1-related proteins family (AF069)
Api g 4	60	1	[61–116]	plant	Profilin family (AF051)
Gad m 1.0101	130	1	[1–41]	fish	EF-hand family (AF007)
Gad m 1.0102	131	1	[1–41]	fish	EF-hand family (AF007)
Gal d 3	6	18	[5–34],[29–51],[33–54],[40–63],[144–177],[266–288],[270–300],[293–316],[302–325],[339–365],[351–375],[427–465],[521–542],[606–631],[616–639],[624–653],[662–689],[673–700]	egg	Transferrin family (AF068)
Gal d 5	8	13	[14–35],[90–111],[92–113],[125–149],[131–160],[191–238],[226–247],[334–357],[399–425],[406–427],[413–437],[440–464],[481–503],[530–568]	egg	Serum albumin family (AF056)
Gly m 3.0101	43	1	[5–45]	plant (leguminous)	Profilin family (AF051)
Gly m 4.0101	44	2	[1–39],[21–70]	plant (leguminous)	
Gly m Bd 60K	41	5	[179–235],[286–372],[216–274],[358–412],[488–583]	plant (leguminous)	Cupin superfamily (AF045)
Hor v 15.0101	25	6	[24–86],[68–119],[103–126],[109–130],[119–140],[124–146]	plant (cereal)	Prolamin superfamily (AF050)
Ory s 1.0101	37	4	[137–187],[57–99],[168–206],[205–236]	plant (cereal)	Expansin, C-terminal domain family (AF093)
Ory s 12.0101	38	2	[9–79],[63–113]	plant (cereal)	Profilin family (AF051)
Ran e 1.0101	33	3	[1–22],[32–53],[64–85]	amphibian	EF-hand family (AF007)
Sal s 1.0101	22	3	[9–30],[13–34],[16–41]	fish	EF-hand family (AF007)
Sal s 1.0201	23	1	[2–25]	fish	EF-hand family (AF007)
Ses i 6.0101	56	10	[25–62],[51–94],[86–128],[112–144],[162–211],[207–243],[231–276],[265–292],[274–323],[393–442]	plant	Cupin superfamily (AF045)
Tri a 3	36	4	[1–51],[66–102],[32–67],[92–118]	plant (cereal)	Expansin, C-terminal domain family (AF093)
Tri a TPI	29	2	[12–47],[32–64]	plant (cereal)	Triosephosphate isomerase family (AF032)
Zea m 14.0101	39	3	[7–51],[50–98],[84–120]	plant (cereal)	Prolamin superfamily (AF050)
Total number of potential epitopes		83			

[1] Allergen names according to WHO-IUIS and Allergome database.

[2] ID in the BIOPEP Database of allergenic proteins

of allergens in the AllFam database is based on the presence of domains summarised in the Pfam database [Radauer *et al.*, 2008]. The presence of characteristic domains is a common basis of protein function annotation [Dessailly *et al.*, 2009], although the alternative solutions of this task are recently proposed [Petrey & Honig, 2009]. Analysis of bioactive peptide profiles of protein containing various domains has been published by Iwaniak & Dziuba [2009]. Information about affiliation of individual allergens to families is available (apart from AllFam database) in Allergome and BIOPEP databases. Families of α/β-caseins and κ-caseins reveal high A values and PPV 100% (with one exception). The PPV of lipocalins (β-lactoglobulins) strongly depends on A value. It indicates that for some proteins belonging to this family knowledge concerning experimental epitopes may be incomplete. Bovine β-lactoglobulin (BIOPEP ID 14) was extensively studied to find fragments interacting with the immunological system. Other proteins from this family contain epitopes identical with these from bovine protein. Lipocalins with a low number of such epitopes are also characterised by a low PPV value. Prediction of epitope location using EVALLER program appears to be promising for proteins revealing low A value and high PPV value, such as proteins from tropomyosin family and papain-like cystein protease family. The group with the lowest A (A<0.2) and PPV=100% contains also single proteins from four other families (Table 4). Among proteins with a low number of known sequential epitopes prediction quality is rather poor for cupin superfamily. This family does not contain proteins with PPV=100%. Four of the six proteins from this family reveal PPV below 50%. The last family is EF--hand family (fish and amphibian parvalbumins). Among these proteins one has revealed PPV=100%, but two – PPV=0. Parvalbumins are characterised by high sequence variability. Their predicted secondary structure may also vary between species as shown on the example of proteins from *Cyprinus carpio* and *Salmo salar* [Iwaniak & Dziuba, 2011]. It is difficult to take any conclusions about predictability of epitope occurrence in proteins form the EF-hand family.

Allergens from the BIOPEP database, containing potential epitopes found by EVALLER program but not containing known experimental sequential epitopes are summarised in Table 5. There is also no data concerning discontinuous epitopes or data concerning these allergens in IEDB or IEDB-3D [Ponomarenko *et al.*, 2011], although we cannot exclude finding such epitopes in the future. Most of allergens indicated in Table 5 are of plant origin (cereals or leguminous plants). Some of allergen AllFam families indicated in Table 5 are represented also among the ones containing both potential and experimental epitopes (Table 4). Proteins from the serum albumin, prolamine or Bet v 1-related family were characterised by a high positive predictive value with a relatively low frequency of epitope occurrence. Results concerning these families look promising in contrast with the proteins belonging to the Cupin and EF-hand families. On the other hand, we cannot exclude that these families contain proteins characterised by poor predictive values such as members of the EF-hand family with PPV=0. The BIOPEP database does not contain proteins from the Profilin, Transferrin and Expansin C-terminal Domain families containing both experimental and potential epitopes.

A comparison with the experimental results is the only way to evaluate *in silico* prediction. Fragments typical of allergenic proteins, generated by program EVALLER reveal usually good (*ca.* 80%) likelihood of coverage with experimental epitopes. The suggestion that they cover such epitopes [Björklund *et al.*, 2005] is thus generally confirmed for the used dataset of food allergenic proteins and their epitopes. The positive predictive value may, however, vary among allergen families. The possibility of continuous update is a characteristic property of bioinformatic tools and is considered as major advantage [Wren & Bateman, 2008]. Further enrichment of both databases: BIOPEP and database used by EVALLER program, may lead to the improvement of the Positive Predictive Value concerning overlapping between potential and experimental epitopes.

CONCLUSIONS

Program EVALLER generates fragments characteristic for allergenic proteins (so called Filtered Length-Adjusted Allergenic Peptides – FLAPs). Most of such peptides present in the BIOPEP database of allergenic proteins (*ca.* 80 %) fully or partially overlap with the known experimental ones. Program EVALLER may thus be used as a tool for predicting epitope location although its potential varies considerably among allergen families. The predictive potential was good for milk proteins (α/β-caseins family and κ-caseins family), although these proteins possess numerous experimental epitopes. The predictive potential appears good also for invertebrate tropomyosins (tropomyosins family) and poor for plant allergens from the Cupin superfamily.

ACKNOWLEDGEMENTS

This work has been financed from grant No. NN 312 286 333 of the Ministry of Science and Higher Education.

REFERENCES

1. Altschul S.F., Madden T.L., Schäffer A.A., Zhang J., Zhang Z., Miller W., Lipman D.J., Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res., 1997, 25, 3389–3402.
2. Ayuso R., Lehrer S.B., Reese G., Identification of continuous, allergenic regions of the major shrimp allergen Pen a 1 (tropomyosin). Int. Arch. Allergy Immunol., 2002, 127, 27–37.
3. Björklund Å., Soeria-Atmadja D., Zorzet A., Hammerling U., Gustafsson M.G., Supervised identification of allergen-representative peptides for *in silico* detection of potentially allergenic proteins. Bioinformatics, 2005, 21, 39–50.
4. Bohle B., T-cell epitopes of food allergens. Clin. Rev. Allergy Immunol., 2006, 30, 97–108.
5. Cianferoni A., Spergel J.M., Food allergy: review, classification and diagnosis. Allergol. Int., 2009, 58, 457–466.
6. Cummings A.J., Knibb R.C., King R.M., Lucas J.S., The psychosocial impact of food allergy and food hypersensitivity in children, adolescents and their families: a review. Allergy, 2010, 65, 933–945.
7. Darewicz M., Dziuba B., Minkiewicz P., Dziuba J., The preventive potential of milk and colostrum proteins and protein fragments. Food Rev. Int., 2011, 27, 357–388.
8. Dessailly B.H., Redfern O.C., Cuff A., Orengo C.A., Exploiting structural classifications for function prediction: towards a domain grammar for protein function. Curr. Opin. Struct. Biol., 2009, 19, 349–356.
9. Dziuba J., Iwaniak A., Minkiewicz P., Computer-aided characteristics of proteins as potential precursors of bioactive peptides. Polimery, 2003, 48, 50–53.
10. Gendel S.M., Allergen databases and allergen semantics. Regulatory Toxicol. Pharmacol., 2009, 54(3 Suppl.), S7-S10.
11. Goodman R.E., Practical and predictive bioinformatic methods for the identification of potentially cross-reactive protein matches. Mol. Nutr. Food Res., 2006, 50, 655–660.
12. Gowthaman U., Agrewala J.N., *In silico* methods for predicting T-cell epitopes: Dr Jekyll or Mr Hyde? Expert Rev. Proteom., 2009, 6, 527–537.
13. Ishikawa M., Nagashima Y., Shiomi K., Identification of the oyster allergen Cra g 2 as tropomyosin. Fisheries Sci., 1998a, 64, 854–855.
14. Ishikawa M., Ishida M., Shimakura K., Nagashima Y., Shiomi K., Purification and IgE-binding epitopes of a major allergen in the gastropod *Turbo cornutus*. Biosci. Biotech. Biochem., 1998b, 62, 1337–1343.

15. Ishikawa M., Suzuki F., Ishida M., Nagashima Y., Shiomi K., Identification of tropomyosin as a major allergen in the octopus *Octopus vulgaris* and elucidation of its IgE-binding epitopes. Fisheries Sci., 2001, 67, 934–942.
16. Iwaniak A., Dziuba J., Analysis of domains in selected plant and animal food proteins – precursors of biologically active peptides. Food Sci. Technol. Int., 2009, 15, 179–191.
17. Iwaniak A., Dziuba J., BIOPEP-PBIL tool for analysis of the structure of biologically active motifs derived from food proteins. Food Technol. Biotechnol., 2011, 49, 118–127.
18. Jain E., Bairoch A., Duvaud S., Phan I., Redaschi N., Suzek B.E., Martin M.J., McGarvey P., Gasteiger E., Infrastructure for the life sciences: design and implementation of the UniProt website. BMC Bioinform., 2009, 10, Article No 136.
19. Jędrychowski L., Wróblewska B., Szymkiewicz A., State of the art on food allergens – a review. Pol. J. Food Nutr. Sci., 2008, 58, 165–175.
20. Kim J.S., Sicherer S., Should avoidance of foods be strict in prevention and treatment of food allergy? Curr. Opin. Allergy Clin. Immunol., 2010, 10, 252–257.
21. Mari A., Scala E., Palazzo P., Ridolfi S., Zennaro D., Carabella G., Bioinformatics applied to allergy: Allergen databases, from collecting sequence information to data integration. The Allergome platform as a model. Cell. Immunol., 2006, 244, 97–100.
22. Mari A., Rasi C., Palazzo P., Scala E., Allergen databases: current status and perspectives. Curr. Allergy Asthma Rep., 2009, 9, 376–383.
23. Martinez Barrio A., Soeria-Atmadja D., Nistér A., Gustafsson M.G., Hammerling U., Bongcam-Rudloff E., EVALLER: a web server for *in silico* assessment of potential protein allergenicity. Nucleic Acids Res., 2007, 35, W694-W700.
24. Minkiewicz P., Dziuba J., Gładkowska-Balewicz I., Update of the list of allergenic proteins from milk, based on local amino acid sequence identity with known epitopes from bovine milk proteins – a short report. Pol. J. Food Nutr. Sci., 2011, 61, 153–158.
25. Monaci L., Tregoat V., van Hengel A.J., Anklam E., Milk allergens; their characteristics and their detection in food: a review. Eur. Food Res. Technol., 2006, 223, 149–179.
26. Pearson W.R., Searching protein sequence libraries: comparison of the sensitivity and selectivity of the Smith-Waterman and FASTA algorithm. Genomics, 1991, 11, 635–650.
27. Pearson W.R., Flexible sequence similarity searching with the FASTA3 program package. Methods Mol. Biol., 2000, 132, 185–219.
28. Petrey D., Honig B., Is protein classification necessary? Toward alternative approaches to function annotation. Curr. Opin. Struct. Biol., 2009, 19, 363–368.
29. Pomés A., Relevant B cell epitopes in allergic disease. Int. Arch. Allergy Immunol., 2010, 152, 1–11.
30. Ponomarenko J., Papangelopoulos N., Zajonc D.M., Peters B., Sette A., Bourne P.E., IEDB-3D: structural data within the immune epitope database. Nucleic Acids Res., 2011, 39, D1164-D1170.
31. Prescott S.L., Bouygue G.R., Videky D., Fiocchi A., Avoidance or exposure to foods in prevention and treatment of food allergy? Curr. Opin. Allergy Clin. Immunol., 2010, 10, 258–266.
32. Pulido A., Ruisánchez I., Boqué R., Rius F.X., Uncertainty of results in routine qualitative analysis. Trends Anal. Chem., 2003, 22, 647–654.
33. Radauer C., Bublin M., Wagner S., Mari A., Breiteneder H., Allergens are distributed into few protein families and possess a restricted number of biochemical functions. J. Allergy Clin. Immunol., 2008, 121, 847–852.
34. Reese G., Ayuso R., Carle T., Lehrer S.B., IgE-binding epitopes of shrimp tropomyosin, the major allergen Pen a 1. Int. Arch. Allergy Immunol., 1999, 118, 300–301.
35. Reese G., Ayuso R., Leong-Kee S.M., Plante M., Lehrer S.B., The IgE-binding regions of the major allergen Pen a 1: Multiple epitopes or intramolecular cross-reactivity? Int. Arch. Allergy Immunol., 2001, 124, 103–106.
36. Reese G., Viebranz J., Leong-Kee S.M., Plante M., Lauer I., Randow S., Moncin M.S., Ayuso R., Lehrer S.B., Vieths S., Reduced allergenic potency of VR9–1, a mutant of the major shrimp allergen Pen a 1 (tropomyosin). J. Immunol., 2005, 175, 8354–8364.
37. Salimi N., Fleri W., Peters B., Sette A., Design and utilization of epitope-based databases and predictive tools. Immunogenetics, 2010, 62, 185–196.
38. Sammut S.J., Finn R.D., Bateman A., Pfam 10 years on: 10 000 families and still growing. Brief. Bioinform., 2008, 9, 210–219.
39. Shanti K.N., Martin B.M., Nagpal S., Metcalfe D.D., Rao P.V., Identification of tropomyosin as the major shrimp allergen and characterization of its IgE-binding epitopes, J. Immunol., 1993, 151, 5354–5363.
40. Skripak J.M., Sampson H.A., Towards a cure for food allergy. Curr. Opin. Immunol., 2008, 20, 690–696.
41. Smith T.F., Waterman M.S., Identification of common molecular subsequences. J. Mol. Biol., 1981, 147, 195–197.
42. Soeria-Atmadja D., Lundell T., Gustafsson M.G., Hammerling U., Computational detection of allergenic proteins attains a new level of accuracy with *in silico* variable-length peptide extraction and machine learning. Nucleic Acids Res., 2006, 34, 3779–3793.
43. Steckelbroeck S., Ballmer-Weber B.K., Vieths S., Potential, pitfalls and prospects of food allergy diagnostics with recombinant allergens or synthetic sequential epitopes. J. Allergy Clin. Immunol., 2008, 121, 1323–1330.
44. The UniProt Consortium, Ongoing and future developments at the Universal Protein Resource. Nucleic Acids Res., 2011, 39, D214-D219.
45. Tomar N., De R. K., Immunoinformatics: An integrated scenario. Immunology, 2010, 131, 153–168.
46. Tong J.C., Ren E.C., Immunoinformatics: current trends and future directions. Drug Discov. Today, 2009, 14, 684–689.
47. Vaughan K., Greenbaum J., Kim Y., Vita R., Chung J., Peters B., Broide D., Goodman R., Grey H., Sette A., Towards defining molecular determinants recognized by adaptative immunity in allergic disease: an inventory of the available data. J. Allergy, 2010, Article No 628026.
48. Vita R., Zarebski L., Greenbaum J.A., Emami H., Hoof I., Salimi N., Damle R., Sette A., Peters B., The Immune Epitope Database 2.0. Nucleic Acids Res., 2010, 38, D854-D862.
49. Wren J.D., Bateman A., Databases, data tombs and dust in the wind. Bioinformatics, 2008, 24, 2127–2128.

Received May 2011. Revision received October and accepted November 2011. Published on-line on the 28th of March 2012.