Acta Sci. Pol., Technol. Aliment. 12(1) 2013, 101-112
pISSN 1644-0730 eISSN 1889-9594 www.food.actapol.net/

PEPTIDES, SPECIFIC PROTEOLYSIS PRODUCTS, AS MOLECULAR MARKERS OF ALLERGENIC PROTEINS – *IN SILICO* STUDIES*

Marta Dziuba✉, Piotr Minkiewicz, Marianna Dąbek

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

ABSTRACT

The objective of this study was to analyse allergenic proteins by identifying their molecular biomarkers for detection in food using bioinformatics tools. The protein and epitope sequences were from BIOPEP database, proteolysis was simulated using BIOPEP program and UniProt database screening via BLAST and FASTA programs. The biomarkers of food proteins were proposed: for example for whey proteins – TPEVDDEALEKFDKALKALPMHIR (β-Lg: fragment 141-164), chicken egg – AAVSVDCSEYPKPDCTAEDRPL (ovomucoid: 156-177), wheat – KCNGTVEQVESIVNTLNAGQIASTDVVEVVVSPPY (triose phosphate isomerase: 12-46) and peanuts – QARQLKNNNPFKFFVPPFQQSPRAVA (arachin: 505-530). The results are annotated in the BIOPEP database of allergenic proteins and epitopes, available at http://www.uwm.edu.pl/biochemia. The epitope-receptor interactions are attributed to the epitope's sequence and suggest that *in silico* proteolysis products showing the highest degree of sequence identity with an epitope or its part are characteristic of a given protein or a group of cross-reactive homologs. The protein markers from basic food groups were proposed based on the above assumption.

Key words: allergens, biomarkers, database of allergenic proteins, proteolysis simulation, sequence analysis, *in silico* analysis

INTRODUCTION

Allergy is a widespread health problem that affects a growing number of people world-wide. The correct identification of allergens supports the correct determination of their biochemical and immunological parameters, and it contributes to an explanation of the allergy mechanism. Food allergens are most often identified by immunochemical methods, including immunoblotting, immunoelectrophoresis and ELISA [Besler 2001]. Allergenic proteins may also be identified by mass spectrometry [Fæste et al. 2011, Picariello et al. 2011]. The most effective strategy for detecting and identifying allergenic proteins with mass spectrometry involves specific hydrolysis products known as biomarkers [Picariello et al. 2011, Cuollo et al. 2010, Monaci et al. 2010, Al-Shahib et al. 2010, Ansari et al. 2011].

Research into allergic reactions to specific food proteins requires biopsy examinations in human subjects (before and after the consumption of foods that produce pathophysiological symptoms). Studies of the

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

✉dziuba@uwm.edu.pl

type are costly, they require a "reasonable" number of volunteers, effective diagnosis, and they raise moral objections. For this reason, papers that propose markers for the classification of potentially allergenic food proteins are of great value. Computer-assisted techniques are increasingly often deployed to identify the properties of proteins and peptides [Minkiewicz et al. 2008, Dziuba and Dziuba 2010], including immunological and allergenic characteristics [Korber et al. 2006, Tomar and De 2010]. The BIOPEP database of proteins and bioactive peptides [Minkiewicz et al. 2008, Dziuba and Iwaniak 2006] which is available on the website of the Chair of Food Biochemistry at the University of Warmia and Mazury in Olsztyn is such a tool. The database contains the amino acid sequences of proteins and bioactive peptides, it features an application for designing the release of bioactive peptides from their precursors during proteolysis, as well as reference sources. Database algorithms may be applied to accumulate and process information about protein allergenicity.

The aim of this study was to characterise the products of specific protein proteolysis as molecular biomarkers of the quality and safety of food products and raw materials. Our research involved the following stages: (1) development of a database of allergenic proteins and classification of cereal, fish, egg, soy, milk, peanut, crustacean, sesame and celery proteins as potential precursors of allergenic peptides based on their peptide profiles, (2) *in silico* proteolysis of allergenic proteins by endopeptidases of various specificity, (3) evaluation of similarities between the products of specific protein proteolysis and the epitopes of allergenic proteins, and (4) identification of peptides, specific proteolysis products, as molecular markers of allergenic proteins or protein groups.

MATERIAL AND METHODS

Databases and software

The databases and applications used in this study are listed in Table 1 with the website addresses. The protein amino acid sequences were supplied by UniProt and National Center for Biotechnology Information databases. The majority of allergens are also listed in the database of allergenic proteins related to the UniProt database of protein sequences, as well as Allergome, Allergen Database for Food Safety, Allergen Online, SDAP and WHO-IUIS databases.

Structure of the BIOPEP database of allergenic proteins and their epitopes

The database of allergenic proteins was created with the use of data storage methods and the search engine of the previously described database of proteins and biologically active peptides [Minkiewicz et al. 2008, Dziuba and Dziuba 2010, Dziuba and Iwaniak 2006]. The database of allergenic proteins and their epitopes will ultimately constitute an integral part of BIOPEP. Algorithms for evaluating the potential biological activity of proteins, such as the biological activity profile and the occurrence frequency of bioactive motifs (A) (e.g. an epitope), are determined for all proteins listed in the database, including allergenic proteins. The database of allergenic proteins and their epitopes is described in detail in section "Structure of the allergenic protein database".

In silico proteolysis of allergenic proteins

Allergenic proteins were subjected to *in silico* proteolysis based on BIOPEP procedures. A total of 44 amino acid sequences of allergenic proteins from the database were analysed (this number of allergenic proteins had been entered at the initial stages of database development). Proteolytic processes were simulated by 27 proteolytic enzymes of various specificities collected in BIOPEP using the "single-enzyme hydrolysis" option.

Cross-coverage between specific proteolysis products and epitopes of allergenic proteins

Epitope sequences from the BIOPEP database were analysed. Sequence identity between the expected proteolysis products and the corresponding epitopes was calculated using the below formula:

$$SCC = (n_c/n_e) \times (n_c/n_p) \times 100\% \qquad (1)$$

where:

SCC – sequence cross-coverage between the expected proteolysis product and the corresponding epitope,

n_c – number of amino acid residues shared by the expected proteolysis product and the epitope,

Table 1. Website sources used in the study

Databases and applications	Website address	References
Algpred	http://www.imtech.res.in/raghava/algpred	Saha and Raghava [2006]
Allergen Database for Food Safety	http://allergen.nihs.go.jp/ADFS/	Nakamura et al. [2005]
Allergen Online	http://www.allergenonline.org/	Gendel [2009]
Allergome	http://www.allergome.org/	Mari et al. [2009]
AllFam	http://www.meduniwien.ac.at/allergens/allfam/	Radauer et al. [2008]
UniProt list of allergens	http://www.uniprot.org/docs/allergen	The UniProt Consortium [2012]
BIOPEP	http://www.uwm.edu.pl/biochemia	Dziuba and Iwaniak [2006]
BLAST	http://www.ebi.ac.uk/Tools/blast/	Altschul et al. [1997]
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	http://www.immuneepitoApe.org/	Vita et al. [2010]
FASTA	http://www.ebi.ac.uk/fasta33/	Pearson [2000]
National Center for Biotechnology Information (NCBI)	http://www.ncbi.nlm.nih.gov/guide/proteins/	Sayers et al. [2012]
NCBI Taxonomy database	http://www.ncbi.nlm.nih.gov/taxonomy	Federhen [2012]
SDAP	http://fermi.utmb.edu/SDAP/	Ivanciuc et al. [2009]
UniProt	http://www.uniprot.org	The UniProt Consortium [2012]
WHO-IUIS allergen database	http://www.allergen.org/	

n_e – number of amino acid residues in the epitope sequence,

n_p – number of amino acid residues in the sequence of the proteolytic product.

Distribution of potential markers

WU-BLAST and FASTA applications available on the website of the European Bioinformatics Institute (EBI) were used to analyse the distribution of fragments which are potential markers. The release of a proteolytic product from various proteins by the same enzyme was investigated using the sequences of potential proteolytic products with the preceding and following residues as query sequences which corresponded to the enzyme specificity. The following search parameters were set in the BLAST program: PAM 10 matrix, expected noise level – 1000, number of hits – 500, score sorting from highest to lowest. The remaining parameters were set at default. The search in the FASTA application was based on default parameters.

RESULTS AND DISCUSSION

At the initial stage of the study, a database of allergenic proteins was created and included to BIOPEP. It contains information about the allergenic properties of proteins and presence of epitopes. The new database supports the search for new allergens based on the presence of epitopes shared with the previously identified allergenic proteins. It is supplemented with new sequences of allergenic proteins and new epitopes. Users can contribute information about allergenic proteins and/or epitopes in the "Send your sequence" section on the BIOPEP homepage.

BIOPEP algorithms also allow calculating quantitative parameters of bioactive fragments or epitopes in proteins, performing *in silico* proteolysis using the catalogued enzymes and determining the biological activity of a short sequence entered by the user. The database of allergenic proteins and their epitopes is available as BIOPEP sub-base at http://www.uwm.edu.pl/biochemia.

Structure of the allergenic protein database

At present, the sub-base of allergens contains data on 135 allergenic proteins and the sequences of potential and experimentally determined epitopes. From the existing list of 135 allergenic protein, 60 contain both experimentally derived linear epitopes as well as filtered length-adjusted allergenic peptides (FLAP) determined by the EVALLER program. A comparison between the location of allergenic peptides predicted by EVALLER and the location of experimental epitopes in the sequences of 60 analysed proteins revealed complete or partial overlaps between around 80% of allergenic peptides and experimental epitopes [Minkiewicz et al. 2012]. The fragments of proteins selected by EVALLER will be referred to as potential epitopes in successive parts of this article.

Table 2 presents the data from each section of the database of allergenic proteins and their epitopes. The "Motifs" tab contains the sequences of allergenic proteins, linear epitopes determined experimentally and theoretically predicted (potential) epitopes. The epitopes are presented in a format compatible with the BIOPEP

Table 2. Data regarding food allergenic proteins annotated in the BIOPEP database

Section	Contents
Motifs	Identification number in the BIOPEP database, name of the protein, name according to the nomenclature of allergens; amino acid sequence; number of amino acid residues, molecular mass, sequences of linear epitopes determined experimentally or predicted theoretically, in a format compatible with the BIOPEP search engine which supports the creation of epitope profiles, occurrence frequency of epitopes in protein sequence.
References	Bibliographic data on a paper describing protein sequences or information about protein sequences entered into the UniProt database (in most cases) or the NCBI database.
Additional information	Brief information on proteins contained in the UniProt database or the NCBI database; references regarding allergenicity (for proteins whose experimentally determined linear epitopes are not known), sequences and locations of experimentally determined linear epitopes (if present in a given protein), identification numbers of epitopes in the IEDB-AR database (for epitopes present in this database), references regarding experimentally determined linear epitopes, sequences and locations of theoretically predicted linear epitopes (if present in a given protein), references regarding programs used for theoretical prediction of epitopes (Algpred and/or EVALLER).
Homology	Information about attachment in allergenic protein families in the AllFam database. Information about proteins that share experimentally determined linear epitopes with a given protein.
Biomarkers	Potential biomarkers selected *in silico*, divided into three categories: markers of precursor, markers of group of proteins and pepsin markers of group of proteins. Data on potential biomarkers: sequences, location in the sequences of an allergenic proteins, sequence cross-coverage with an epitopes, enzymes releasing a markers, data on proteins or a groups of proteins which may be detected using a biomarkers.
Database reference	Entry name and number in the UniProt and/or NCBI databases, species identifier according to the NCBI taxonomy database, information on the presence of analyzed proteins in other databases of allergenic proteins (in most cases Allergome and WHO-IUIS).

search engine which supports the creation of epitope profiles, described below. Epitope sequences are presented in the following order: experimental epitopes followed by theoretically predicted (potential) epitopes both from the N-terminus to the C-terminus of the protein chain. The "Motifs" tab lists proteins according to UniProt and/or NCBI nomenclature as well as proteins termed in accordance with the principles of allergen nomenclature [King et al. 1995] applied in the WHO-IUIS database, including the isoallergen number. If allergens are listed under different names in various databases, the name used in the Allergome database is given, and alternative names are provided in the "Database reference" tab. The occurrence frequency of epitopes in the protein sequence, defined as the ratio of the number of epitopes (experimental and potential, predicted *in silico*) to the number of amino acid residues in the protein sequence, marked with the letter A in the "Motifs" tab, similarly to the corresponding parameter calculated for bioactive fragments [Dziuba and Iwaniak 2006]. Parameter A is calculated automatically when the protein sequence and epitope location data are entered. The "Additional information" tab contains epitope sequences, ordered identically, but presented in a more user-friendly form than the "Motifs" section. Shared experimentally determined linear epitopes in various proteins are shown in the "Homology" section. It describes the epitope sequence, its location in the protein chain as well as other proteins containing that epitope. So far, all proteins containing epitopes shared with given protein have been recognised as its homologs. Proteins containing shared epitopes are found in bovine milk (*Bos taurus*) and in the milk of other species [Minkiewicz et al. 2011]. Another example are tropomyosins of crustaceans and other invertebrates. All invertebrate tropomyosins listed in the BIOPEP database (ID 3, 24, 62, 63, 64, 65, 72, 73 and 93) contain epitopes shared with allergen Pen a 1.0102 (BIOPEP ID 76) which was subjected to experimental epitope mapping. The "Biomarkers" tab contains information about peptides that can be used as biomarkers of various allergens. Classification of biomarkers is discussed in section "*In silico* classification of potential markers of allergenic proteins and main groups of allergenic food proteins".

The proposed sub-base features tools for sequence analysis similarly to the BIOPEP database. A diagram illustrating analytical options for protein and peptide sequences in the BIOPEP database is presented in Figure 1. In the "Analysis" tab, the first option supports the creation of the profiles of potential biological activity [Minkiewicz et al. 2008, Dziuba and Dziuba 2010, Dziuba and Iwaniak 2006] and epitope locations.

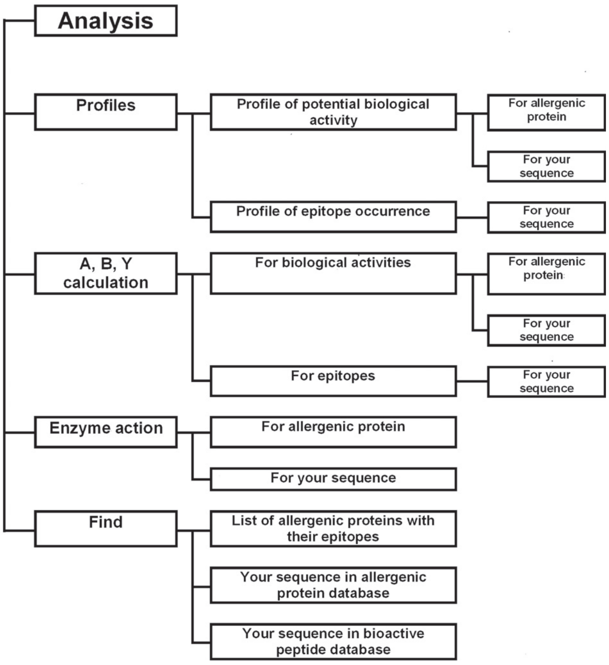
Analysis
Profiles
Profile of potential biological activity
For allergenic protein
For your sequence
Profile of epitope occurrence
For your sequence
A, B, Y calculation
For biological activities
For allergenic protein
For your sequence
For epitopes
For your sequence
Enzyme action
For allergenic protein
For your sequence
Find
List of allergenic proteins with their epitopes
Your sequence in allergenic protein database
Your sequence in bioactive peptide database

Fig. 1. A scheme illustrating sequence analysis using the options available in the database of allergenic proteins and their epitopes

The determination of epitope profiles for protein sequences entered by the user supplies information about the location of epitopes shared with previously identified allergens. This option differs from the algorithms used previously to search for common epitopes based on similarities between protein sequences [Kleter and Peijnenburg 2002] or involving the use of pentapeptides which are epitope fragments as query sequences [Kanduc 2008] or the use of epitope sequences [Minkiewicz et al. 2011]. The sequences in the BIOPEP database may also be analysed with the use of other bioinformatics tools available in the "Useful links" tab.

In silico classification of potential markers of allergenic proteins and main groups of allergenic food proteins

In silico proteolysis were carried out on 44 allergenic proteins of cereals, fish, eggs, soybean, milk, peanuts, crustaceans, sesame and celery, based on "Enzyme action" option. In line with the EVALLER algorithm [Martinez Barrio et al. 2007], it was assumed that the product of *in silico* proteolysis of an allergenic protein, whose structure is identical to that of the epitope predicted in the above application or its significant part, is characteristic of that protein. Biomarkers of the analysed allergenic proteins were identified, and classified as unique when observed in single proteins or as marker of protein group when observed in at least two homologous, cross-reactive allergenic proteins. Exemplary marker categories data are shown in Table 3. Comprehensive information about markers of precursors can be found in the "Biomarkers" tab of BIOPEP database. An analysis performed using BLAST and FASTA applications revealed that the identified fragments were present in 60 proteins listed in BIOPEP. In the group of 44 identified fragments, 33 were present in single allergens, 10 could be used as markers for several (2 to 10) allergens, whereas the presence of 1 fragment was noted in more than 10 proteins listed in BIOPEP. In the UniProt database, 16 fragments were present in single proteins, 22 fragments were potential markers for several (2-10) proteins, whereas 5 fragments could be markers for 11 to 100 proteins. A single fragment was observed in 118 proteins. The most common fragment, AGDSDGDGK, was a marker of fish and amphibian parvalbumins and proteins, and its presence (including possible release by thermolysin – EC 3.4.24.27) was observed in 118 proteins, 12 of which are described in the BIOPEP database. The above fragment was identified by analysing the Cyp c 1.01 carp parvalbumin sequence (BIOPEP ID 18). Allergen Cyp c 1 may be used as a model compound in diagnosing allergic reactions to parvalbumin of other fish species [Agabriel et al. 2010].

A single peptide characterised by the highest cross-coverage with the corresponding epitope was selected in each of the analysed groups of food products and raw materials (Table 4). Enzymes that specifically released peptides, the proposed biomarkers of allergenic proteins, are also listed in the above table. Trypsin and pepsin are the most popularly used in allergen research [Fæste et al. 2011, Schnell and Herman 2009, Wickham et al. 2009], but peptides released by other enzymes can also be used. Fragments liberated by enzymes other than trypsin, which are most similar to fragments of allergenic proteins and which reveal maximum difference from fragments of non-allergenic proteins [Martinez Barrio et al. 2007], can also serve as markers. The results of computer-simulated peptic proteolysis of proteins from ten most allergic food groups are presented in Table 5. In most cases, the degree of overlap between epitopes and pepsin-released

Table 3. Exemplary information about particular categories of biomarkers

Marker category	Marker data
Marker of precursor (in that case Ara h 1; BIOPEP ID 45)	ELHLLGFGINA (precursor fragment 520-530). Sequence cross-coverage with epitope 91%. Fragment may be released by leukocyte elastase EC 3.4.21.37. Fragment occurs in 8 sequences of glycinins and conarachins from peanut (*Arachis hypogaea*) annotated as an allergen Ara h 1 (Allergome code 50). Potential marker of allergen Ara h 1 and peanut proteins.
Marker of protein group (in that case peanut proteins)	QARQLKNNNPFKFFVPPFQQSPRAVA (Ara h 4.0101 (ID 60) fragment: 505-530). Sequence cross-coverage with epitope 100%. Fragment may be released by V-8 protease (glutamyl endopeptidase) EC 3.4.21.19.
Pepsin marker of protein group (in that case peanut proteins)	RGRAHVQVVDSNGNRVYDEEL (Ara h 4.0101 (ID 60) fragment: 421-441). Sequence cross-coverage with epitope 81%. Fragment may be released by pepsin (pH 1.3) EC 3.4.23.1.

Table 4. Proposed biomarkers of food proteins

Proteins	Enzymes	Released peptides	Sequence cross-coverage with epitope, %	Distribution of proposed markers in protein sequences
Whey proteins	Clostripain EC 3.4.22.8	TPEVDDEALEKFDKALKALPMHIR (fr. of β-lactoglobulin 141-164)	89	Fragment present in 6 sequences of bovine (*Bos taurus*) and buffalo (*Bubalus bubalis*) b-lactoglobulin, including Bos d 5 (BIOPEP ID 14) and Bub b BLG (BIOPEP ID 108). A potential marker of bovine and buffalo β-lactoglobulin and whey proteins.
Bovine casein	Calpain EC 3.4.22.17	QHQKAMKPWIQPKTKVIPYVRYL (fr. of α_{s2}-casein 200-222)	96	Fragment present only in the sequence of bovine α_{s2}-casein (BIOPEP ID 10; Bos d 8 alpha s2). A potential marker of bovine α_{s2}-casein Bos d 8 alpha s2 and bovine casein.
Chicken egg	Pepsin (pH 1.3) EC 3.4.23.1	AAVSVDCSEYPKPDCTAEDRPL (fr. of ovomucoid 156-177)	96	Fragment present in 2 sequences of chicken (*Gallus gallus*) egg ovomucoids, including Gal d 1.0101 (BIOPEP ID 5). A potential marker of chicken egg ovomucoid Gal d 1.0101 and chicken egg proteins.
Sesame	Prolyl oligopeptidase EC 3.4.21.26	YVFEDQHFITGFRTQHGRMRVLQKFTDRSELLRGIENYRVAILEAEP (fr. of globulin 7S 196-242)	94	Fragment present only in the sequence of sesame (*Sesamum indicum*) globulin 7S, Ses and 3.0101 (BIOPEP ID 53). A potential marker of sesame globulin 7S and sesame proteins.
Soybean	Proteinase P1 (lactocepin) EC 3.4.21.96	EHGRVYHNHEEEAKRLEIFKNN (fr. of allergen Gly m Bd 30K 150-171)	96	Fragment present only in the sequence of allergen Gly m Bd 30K 1 (BIOPEP ID 42). A potential marker of allergen Gly m Bd 30K 1 and soybean proteins.
Peanuts	V-8 protease (Glutamyl endopeptidase) EC 3.4.21.19	QARQLKNNNPFKFFVPPFQQSPRAVA (fr. of allergen Ara h 4.0101 505-530)	100	Fragment present in the sequences of peanut (*Arachis hypogaea*) arachins Ara h 4.0101 (BIOPEP ID 48) and Ara h Arachin 6 (BIOPEP ID 51). A potential marker of peanut arachins and peanut proteins.
Fish and amphibians	Thermolysin EC 3.4.24.27	AGDSDGDGK (fr. of carp parvalbumin Cyp c 1.01 89-97)	100	Fragment present in 118 sequences of fish and amphibian parvalbumins, including BIOPEP ID: 1, 16, 17, 18, 19, 20, 21, 90, 91, 92, 94 and 96. A potential marker of fish and amphibian parvalbumins and proteins.
Crustaceans	Glycyl endopeptidase EC 3.4.22.25	ESKIVELEEELRVVG (fr. of crab tropomyosin (*Charybdis feriatus*) 187-201)	100	Fragment present in 73 sequences of tropomyosins, including BIOPED ID 62, 63, 72, 73, 76 and 93. A potential marker of crustacean tropomyosins and proteins.
Wheat	Chymase EC 3.4.21.39	KCNGTVEQVESIVNTLNAGQIASTDVVEVVVSPPY (fr. 12-46)	97	Fragment present only in the sequence of wheat (*Triticum aestivum*) triose phosphate isomerase, Tri a TPI (BIOPEP ID 29). A potential marker of wheat allergen Tri a TPI and wheat proteins.
Celery	Calpain EC 3.4.22.17	LGGAKYMVIQGEPNAVIRGKKGSGGVTIKKTGQALVFGVY (fr. 70-109)	71	Fragment present only in the sequence of allergen Api g 4 (BIOPEP ID 60). A potential marker of celery allergen Api g 4 and celery proteins.

Table 5. Peptide markers of most allergenic food groups released by pepsin EC 3.4.23.1 (pH 1.3)

Protein source	Protein name and BIOPEP ID	Protein fragment	Sequence cross-coverage with epitope, %	Distribution of proposed markers in protein sequences
1	2	3	4	5
Chicken egg	Ovotransferrin (conalbumin), precursor, chicken (*Gallus gallus*), allergen Gal d 3 (ID 6)	CQGSGGIPPEKCVASSHEKYF (fr. 521-541)	95	Fragment present in 4 sequences of chicken egg ovotransferrin, including Gal d 3 (BIOPEP ID 6). A potential marker of chicken egg ovotransferrin Gal d 3 and chicken egg proteins.
Casein	Kappa-casein gen. var. A, precursor, bovine (*Bos taurus*), Allergen Bos d 8 kappa (ID12)	GAQEQNQEQPIRCEKDERF (fr. 20-38)	86	Fragment present only in the sequence of bovine κ-casein, Bos d 8 kappa (BIOPEP ID 12). A potential marker of bovine κ-casein and casein.
Whey proteins	Beta-lactoglobulin, gen. var. A, precursor, bovine (*Bos taurus*), Allergen Bos d 5 (ID 14)	VRTPEVDDEAL (fr. 139-149)	48	Fragment present in 5 sequences of bovine (*Bos taurus*) and buffalo (*Bubalus bubalis*) β-lactoglobulins, including Bos d 5 (BIOPEP ID 14) and Bub b BLG (BIOPEP ID 108). A potential marker of bovine and buffalo β-lactoglobulins and bovine or buffalo whey proteins.
Fish	Parvalbumin-beta 1, Atlantic cod (*Gadus morhua*), Allergen Gad m 1.0202 (ID 16)	NDADITAAL (8-16)	41	Fragment present in 17 sequences of parvalbumins, including Gad m 1.0202 (BIOPEP ID 16), Gad m 1.0201 (BIOPEP ID 92), Cyp c 1.01; BIOPEP ID 18) and Cyp c 1.02 (BIOPEP ID 19). A potential marker of Atlantic cod (*Gadus morhua*) parvalbumin Gad m 1, carp (*Cyprinus carpio*) parvalbumin, and Atlantic cod and carp proteins.
Wheat	Agglutinin isolectin 3, fragment, wheat (*Triticum aestivum*), Allergen Tri a agglutinin 3 (ID 30)	CCSQWGYCGL (fr. 103-112)	60	Fragment present in the sequences of wheat (*Triticum aestivum*) agglutinin Tri a agglutinin 3 (BIOPEP ID 30) and barley (*Hordeum vulgare*) agglutinin. A potential marker of wheat and barley agglutinins and proteins.
	Agglutinin isolectin 2, precursor, wheat (*Triticum aestivum*), Allergen Tri a agglutinin 2 (ID 31)	CCSQWGF (fr. 130-136)	50	Fragment characteristic of wheat (*Triticum aestivum*) agglutinins Tri a agglutinin 2 (BIOPEP ID 31) and Tri a 18.0101 (BIOPEP ID 32). A potential marker of wheat agglutinins and wheat proteins.
Soybean	P34 probable thiol protease, soybean (*Glycine max*), allergen Gly m Bd 30K 1 (ID 42)	YTGGIYDGENCTSPYGINHF (fr. 283-302)	78	Fragment present in 2 sequences of soybean (*Glycine max*) proteins including allergen Gly m Bd 30K 1 (BIOPEP ID 42). A potential marker of soybean allergen Gly m Bd 30K 1 and soybean proteins.

Table 5 – cont.

1	2	3	4	5
Peanuts	Conglutin-7, precursor, peanut (*Arachis hypogaea*), Allergen Ara h 2.0101 (ID 46)	AAHASARQQWEL (16-27)	83	Fragment present in 4 conglutins of *Arachis hypogaea* (including allergen Ara h 2.0101 – BIOPEP ID 46), 3 conglutins of *Arachis duranensis* (allergen Ara d 2, Allergome code 5877) and 1 conglutin of *Arachis ipaensis* (Ara and 2.02, Allergome code 5880). A potential marker of allergenic conglutins in plants of the genus *Arachis*.
	Glycinin, peanut (*Arachis hypogaea*), Allergen Ara h 4.0101 (ID 48)	RGRAHVQVVDSNGNRVYDEEL (421-441)	81	Fragment present in 6 sequences of peanut (*Arachis hypogaea*) arachins and glycinins, including Ara h 3.0101 (BIOPEP ID 47), Ara h 4.0101 (BIOPEP ID 48), Ara h Arachin 6 (BIOPEP ID 51). A potential marker of peanut glycinins and arachins and peanut proteins.
Celery	Profilin, celery (*Apium graveolens*), Allergen Api g 4 (ID 60)	GGAKYMVIQGEPNAVIRGKKGSGGVTIKKTGQAL (fr. 71-104)	61	Fragment present only in the sequences of celery (*Apium graveolens*) profilin Api g 4 (BIOPEP ID 60). A potential marker of celery profilin and celery proteins.
Sesame	11S globulin seed storage protein 2, oriental sesame (*Sesamum indicum*), Allergen Ses and 6.0101 (ID 56)	ISIMVPGCAETYQVHRSQRTMERTEASEQQDRGSVRDL (fr. 103-140)	67	Fragment present only in the sequences of sesame (*Sesamum indicum*) globulin 11S Ses and 6.0101 (BIOPEP ID 56). A potential marker of sesame globulin 11S and sesame proteins.
Crustaceans	Tropomyosin, Isoform fast muscle type, Lobster (*Homarus americanus*), Allergen Hom a 1.0102 (ID 63)	VNEKEKYKSIDTEL (fr. 261-274)	80	Fragment present in 29 sequences of tropomyosins, including Hom a 1.0102 (BIOPEP ID 63), Pan s 1 (BIOPEP ID 64), Hom a 1.0101 (BIOPEP ID 72), Met e 1 (BIOPEP ID 73), Pen a 1 (BIOPEP ID 76). A potential marker of crustacean tropomyosins and crustacean proteins.

fragments is lower than for enzymes listed in Table 4. Table 5 presents markers found in the sequence of a single protein (e.g. fragment of bovine k-casein) and homologous proteins of several species (fragments of bovine and buffalo b-lactoglobulin, wheat and barley agglutinin or allergenic conglutins in plants of the genus *Arachis*). The fish parvalbumin fragment present in 17 sequences, including Common carp and Atlantic cod sequences, is most prevalent among pepsin-released markers.

In our follow-up research, peptidomic and proteomic methods are employed to verify the idea of *in silico* determined peptidic markers of allergenic proteins. Correctness of this strategy was confirmed by the results of research [Bucholska 2012, Mogut 2012] concerning peptic hydrolysates of fish and milk proteins analysed by reversed-phase high-performance liquid chromatography with mass spectrometry (RP--HPLC-MS). It was shown that the proposed peptidic biomarkers could be the presence indicator of allergenic protein.

CONCLUSIONS

Research into allergic reactions to specific food proteins requires biopsy examinations in “reasonable” number of human subjects’, effective diagnosis, raises moral objections and is costly. Therefore, works that propose markers allowing classifying and characterising potentially allergenic food proteins are of great value.

Based on the presented protein evaluation criteria with the option of computer-simulated proteolysis in the BIOPEP database, allergenic proteins were analysed by identifying their molecular biomarkers in food materials and processed food products. A database combining information about the allergenic properties of proteins and the presence of epitopes was developed as a sub-base of the existing BIOPEP database. Food proteins were analysed *in silico* to determine the effect of specific proteolysis on the release of allergenic protein epitopes or their fragments. Similarities in the sequence motifs of specific products of protein proteolysis and the known epitopes of allergenic proteins were described.

The results of *in silico* studies were used to identify the molecular biomarkers of allergenic proteins in foods. The above findings can be applied to evaluate the quality and health benefits of foods. The specific biological interactions between an epitope and a receptor can be attributed to the epitope’s chemical structure (amino acid sequence), which suggests that the products of *in silico* proteolysis of allergenic proteins showing the highest degree of sequence identity with an epitope or its part are characteristic of a given protein or a group of homologous proteins that are cross-reactive. The biomarkers of proteins found in basic food groups were identified based on the above assumption. The resulting data are available in the BIOPEP database (http:/www.uwm.edu.pl/biochemia). In our follow-up work, the identified biomarkers and the applied methods may be used to improve existing and develop new techniques for the detection of allergenic proteins in various food groups.

REFERENCES

Agabriel C., Robert P., Bongrand P., Sarles J., Vitte J., 2010. Fish allergy: In Cyp c 1 we trust. Allergy 65, 1483-1484.

Al-Shahib A., Misra R., Ahmod N., Fang M., Shah H., Gharbia S., 2010. Coherent pipeline for biomarker discovery using mass spectrometry and bioinformatics. BMC Bioinformatics 11, art. no. 437.

Altschul S.F., Madden T.L., Schäffer A.A., Zhang J., Zhang Z., Miller W., Lipman D.J., 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25, 3389-3402.

Ansari P., Stoppacher N., Rudolf J., Schuhmacher R., Baumgartner S., 2011. Selection of possible marker peptides for the detection of major ruminant milk proteins in food by liquid chromatography-tandem mass spectrometry. Anal. Bioanal. Chem. 399, 1105-1115.

Besler M., 2001. Determination of allergens in foods. Trends Anal. Chem. 20, 662-672.

Bucholska J., 2012. Identification of markers of potential allergens and bioactive peptides present in fish proteins. Univ. Warmia Mazury Olsztyn, Poland.

Cuollo M., Caira S., Fierro O., Pinto G., Picariello G., Addeo F., 2010. Toward milk speciation through the monitoring of casein proteotypic peptides. Rapid Commun. Mass Spectr. 24, 1687-1696.

Dziuba J., Iwaniak A., 2006. Database of Protein and Bioactive Peptide Sequences. In: Nutraceutical proteins and peptides in health and disease. Eds Y. Mine, F. Shahidi. CRC – Taylor & Francis Boca Raton, 543-563.

Dziuba M., Dziuba B., 2010. *In Silico* analysis of bioactive peptides. In: Bioactive proteins and peptides as functional foods and nutraceuticals. Eds Y. Mine, E.C.Y. Li-Chan, B. Jiang. Blackwell Publ. Oxford, UK, 324-340.

Fæste C.K, Rønnin H.T., Christians U., Granum P.E., 2011. Liquid chromatography and mass spectrometry in food allergen detection. J. Food Protect. 74, 316-345.

Federhen S., 2012. The NCBI taxonomy database. Nucleic Acids Res. 40, 136-143.

Gendel S.M., 2009. Allergen databases and allergen semantics. Regul. Toxicol. Pharmacol. (3 Suppl.) 54, 7-10.

Ivanciuc O., Schein C., Garcia T., Oezguen N., Negi S.S., Braun W., 2009. Structural analysis of linear and conformational epitopes of allergens. Regul. Toxicol. Pharmacol. (3 Suppl.) 54, 11-19.

Kanduc D., 2008. Correlating low-similarity peptide sequences and allergenic epitopes. Curr. Pharmaceut. Design. 14, 289-295.

King T.P., Hoffman D., Løwenstein H., Marsh D.G., Platts-Mills T.A.E., Thomas W., 1995. Allergen nomenclature. J. Aller. Clin. Immunol. 96, 5-14.

Kleter G.A., Peijnenburg A.A., 2002. Screening of transgenic proteins expressed in transgenic food crops for the presence of short amino acid sequences identical to potential, IgE-binding linear epitopes of allergens. BMC Struct. Biol. 2, 8.

Korber B., LaBute M., Yusim K., 2006. Immunoinformatics comes of age. PLoS Comput. Biol. 2, 484-492.

Mari A., Rasi C., Palazzo P., Scala E., 2009. Allergen databases: current status and perspectives. Curr. Aller. Asthma Rep. 9, 376-383.

Martinez Barrio A., Soeria-Atmadja D., Nistér A., Gustafsson M.G., Hammerling U., Bongcam-Rudloff E., 2007. EVALLER: a web server for *in silico* assessment of potential protein allergenicity. Nucleic Acids Res. 35, 694-700.

Minkiewicz P., Dziuba J., Darewicz M., Bucholska J., Mogut D., 2012. Evaluation of *in silico* prediction possibility of potential epitope sequences using experimental data concerning allergenic food proteins summarized in BIOPEP database. Pol. J. Food Nutr. Sci. 62, 151-157.

Minkiewicz P., Dziuba J., Gładkowska-Balewicz I., 2011. 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. 61, 153-158.

Minkiewicz P., Dziuba J., Iwaniak A., Dziuba M., Darewicz M., 2008. BIOPEP database and other programs for processing bioactive peptide sequences. J. AOAC Int. 91, 965-980.

Mogut D., 2012. Identification of peptidic markers of allergenic milk proteins. Univ. Warmia Mazury Olsztyn, Poland [in press].

Monaci L., Losito I., Palmisano F., Visconti A., 2010. Identification of allergenic milk proteins markers in fined white wines by capillary liquid chromatography-electrospray ionization-tandem mass spectrometry. J. Chromatogr. A, 1217, 4300-4305.

Nakamura R., Teshima R., Talagi K., Sawada J.I., 2005. Development of Allergen Database for Food Safety (ADFS): an integrated database to search allergens and predict allergenicity. Bull. Natl. Inst. Health Sci. 123, 32-36.

Pearson W.R., 2000. Flexible sequence similarity searching with the FASTA3 program package. Methods Mol. Biol. 132, 185-219.

Picariello G., Mamone G., Addeo F., Ferranti P., 2011. The frontiers of mass spectrometry-based techniques in food allergenomics. J. Chromatogr. A, 1218, 7386-7398.

Radauer C., Bublin M., Wagner S., Mari A., 2008. Allergens are distributed into few protein families and possess a restricted number of biochemical functions. J. Aller. Clin. Immunol. 121, 847-852.

Saha S., Raghava G.P.S., 2006. AlgPred: prediction of allergenic proteins and mapping of IgE epitopes. Nucleic Acids Res. 34, 202-209.

Sayers E.W., Barrett T., Benson D.A., Bolton E., Bryant S.H., Canese K., et al., 2012. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. 40, 13-25.

Schnell S., Herman R.A., 2009. Should digestion assays be used to estimate persistence of potential allergens in tests for safety of novel food proteins? Clin. Mol. Allergy 7, art. no. 1.

The UniProt Consortium, 2012. Reorganizing the protein space at the Universal Protein Resource (UniProt). Nucleic Acids Res. 40, 71-75.

Tomar N., De R.K., 2010. Immunoinformatics: An integrated scenario. Immunology 131, 153-168.

Vita R., Zarebski L., Greenbaum J.A., Emami H., Hoof I., Salimi N., Damble R., Sette A., Peters B., 2010. The Immune Epitope Database 2.0. Nucleic Acids Res. 38, 854-862.

Wickham M., Faulks R., Mills C., 2009. *In vitro* digestion methods for assessing the effect of food structure on allergen breakdown. Mol. Nutr. Food Res. 53, 952-958.

PEPTYDY, PRODUKTY SPECYFICZNEJ PROTEOLIZY JAKO MOLEKULARNE MARKERY ALERGENNYCH BIAŁEK – BADANIA *IN SILICO*

STRESZCZENIE

Celem przeprowadzonych badań była charakterystyka produktów specyficznej proteolizy białek alergennych jako molekularnych biomarkerów ich obecności w żywności. Wykorzystano sekwencje białek i epitopów znajdujące się w bazie BIOPEP oraz przeprowadzono symulację proteolizy w tej bazie. Użyto programów BLAST i FASTA do przeszukiwania zasobów bazy UniProt. Wytypowano biomarkery białek żywności, między innymi: serwatkowych – TPEVDDEALEKFDKALKALPMHIR (fr. b-Lg 141-164), jaja kurzego – AAVSVDCSEYPKPDCTAEDRPL (fr. owomukoidu 156-177), pszenicy – KCNGTVEQVESI-VNTLNAGQIASTDVVEVVVSPPY (fr. izomerazy fosforanu triozy 12-46) i orzeszków arachidowych – QARQLKNNNPFKFFVPPFQQSPRAVA (fr. arachin 505-530). Wyniki zostały zamieszczone w podbazie alergennych białek i ich epitopów bazy BIOPEP dostępnej na stronie http://www.uwm.edu.pl/biochemia. Specyficzność biologicznego oddziaływania epitopu z receptorem wynika między innymi z określonej chemicznej struktury tego epitopu. Na tej podstawie przyjęto, że produkty proteolizy *in silico* białek alergennych, nakładające się maksymalnie z sekwencjami epitopów, są charakterystyczne dla grup białek alergennych.

Słowa kluczowe: alergeny, biomarkery, baza białek alergennych, symulacja proteolizy, analiza sekwencji, badania *in silico*

Received – Przyjęto: 24.09.2012 Accepted for print – Zaakceptowano do druku: 30.10.2012

For citation – Do cytowania

Dziuba M., Minkiewicz P., Dąbek M., 2013. Peptides, specific proteolysis products, as molecular markers of allergenic proteins – *in silico* studies. Acta Sci. Pol., Technol. Aliment. 12(1), 101-112.