



## Short communication

## Epitopic hexapeptide sequences from Baltic cod parvalbumin beta (allergen Gad c 1) are common in the universal proteome

Piotr Minkiewicz<sup>\*,1</sup>, Justyna Bucholska<sup>1</sup>, Małgorzata Darewicz, Justyna Borawska

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

## ARTICLE INFO

## Article history:

Received 18 July 2012

Received in revised form 14 August 2012

Accepted 14 August 2012

Available online 23 August 2012

## ABSTRACT

The aim of this study was to analyze the distribution of hexapeptide fragments considered as epitopes of Baltic cod parvalbumin beta (allergen Gad c 1) in the universal proteome. Cod (*Gadus morhua* subsp. *callarias*) parvalbumin hexapeptides cataloged in the Immune Epitope Database were used as query sequences. The UniProt database was screened using the WU-BLAST 2 program. The distribution of hexapeptide fragments was investigated in various protein families, classified according to the presence of the appropriate domains, and in proteins of plant, animal and microbial species. Hexapeptides from cod parvalbumin were found in the proteins of plants and animals which are food sources, microorganisms with various applications in food technology and biotechnology, microorganisms which are human symbionts and commensals as well as human pathogens. In the last case possible coverage between epitopes from pathogens and allergens should be avoided during vaccine design.

© 2012 Elsevier Inc. All rights reserved.

## 1. Introduction

Fish is one of the most common food allergens [30,45,51,55]. Parvalbumins play the most significant role in the group of fish allergenic proteins. The most extensively studied allergenic fish protein is parvalbumin beta of the Baltic cod (*Gadus morhua* subsp. *callarias*) known as allergen Gad c 1. The above-mentioned protein was the first allergen to undergo epitope mapping, i.e. analyses of the allergenicity and immunogenicity of various protein fragments [12–17]. Cross-reactivity between parvalbumins from various species was determined experimentally [52,64]. Cross-reactivity also exists between taxonomically unrelated allergens [19]. The identification of fragments containing 6–8 amino acid residues which are identical to a fragment of a known allergen is an official bioinformatics criterion for cataloging a potentially allergenic protein and predict cross-reactivity [4,23]. The identification of fragments characterized by 100% identity and length corresponding to registered epitopes is the simplest criterion for determining allergenic proteins based on in silico predictions [36].

The results of in silico analyses which screened the universal proteome or individual proteomes for fragments similar or identical to oligopeptides used as query sequences have been described by Kleter and Peijnenburg [36], Dummer et al. [11], Darewicz et al. [8], Kanduc [32,33], Kanduc et al. [35], Trost et al. [63] and

Minkiewicz et al. [49]. Synonymous terms of “universal proteome” and “proteomosphere”, which imply a set of all known proteins and proteomes, have been introduced by Kusalik et al. [39] and Shevchenko et al. [60], respectively.

The recent approach to protein classification and function annotation [10,41,54] involves determinations of the appropriate domains, although alternative classification approaches have been proposed [53]. Domain-based classification has been used to identify allergen families [56].

The aim of this study was to analyze the distribution of hexapeptide fragments considered as epitopes of the allergen Gad c 1 in the universal proteome. The distribution of hexapeptide fragments was investigated in various protein families, classified according to the presence of the appropriate domains, and in proteins of plant, animal and microbial species that come into contact with humans.

## 2. Methods

Hexapeptides from the Gad c 1 isoform annotated as Gad c 1.0101 were used as a query sequences. The protein entry name and accession number in UniProt database (<http://www.uniprot.org>) [29,61] are PRVB.GADMC and P02622 respectively. The protein possess code 3290 in Allergome database (<http://www.allergome.org/>) [47,48] and ID No. 17 in the BIOPEP database of allergenic proteins and their epitopes (<http://www.uwm.edu.pl/biochemia>) [49]. The following hexapeptide sequences were used: DEDGFY (ID in the Immune Epitope Database: 7935), IADEDK (ID 25242), KGILSN (ID 30974) and

\* Corresponding author. Tel.: +48 89 523 35 85.

E-mail address: [minkiew@uwm.edu.pl](mailto:minkiew@uwm.edu.pl) (P. Minkiewicz).

<sup>1</sup> These authors contributed equally to this work.

SNADIK (ID 59791). The Immune Epitope Database [65] is available at the website: <http://www.immuneepitope.org/>.

The UniProt database was screened with the use of the WU-BLAST 2 (<http://www.ebi.ac.uk/Tools/blast/>) [2] program with the PAM10 matrix. The following screening parameters were applied: number of scores – 500, number of alignments – 500, expected threshold – 1000. The remaining parameters were set at default values. Only protein fragments with the same length and 100% identity with the query sequence were included in the analysis, in accordance with the recommendations of Minkiewicz et al. [49].

The results of the analysis revealed the location of the studied hexapeptides in the protein sequence. Information about domains within the chain is linked in UniProt records. Only domains covering entire hexapeptide sequences were taken into account. Every domain integrated in the InterPro (<http://www.ebi.ac.uk/interpro/>) [27] database was regarded as a single item. The domains not integrated in InterPro, but present in other domain databases, such as CATH (<http://www.cathdb.info>) [7,66], HAMAP ([www.expasy.org/sprot/hamap](http://www.expasy.org/sprot/hamap)) [42], PANTHER (<http://panther.celera.com>) [62], Pfam B (<http://pfam.sanger.ac.uk/>) [18] or SUPERFAMILY (<http://supfam.org/>) [43], were considered as separate items even if they were attributed to the same function.

The Latin names of species were used as queries in the Scopus literature database (<http://www.scopus.com/home.url>) via the title, keyword and abstract search options. Organisms were classified based on the context of literature records in the database. The following four categories were taken into account for further discussion: animal and plant species relevant for the food industry and/or agriculture, microorganisms considered as useful or potentially useful in food technology, human symbionts and commensals (mainly oral and gastrointestinal microorganisms) and human pathogens and parasites. The AntigenDB database (<http://www.imtech.res.in/raghava/antigendb>) [3] was screened for the information concerning human pathogens. Information about microflora of the human gastrointestinal tract was also obtained from the website of Human Gut Microbiome Project ([http://genome.wustl.edu/genomes/human\\_gut\\_microbiome\\_genomes](http://genome.wustl.edu/genomes/human_gut_microbiome_genomes)) [24]. Those categories were selected due to their possible contact with humans.

All databases were accessed between May and July 2011.

### 3. Results

The number of proteins and domains/families containing individual hexapeptides is presented in Table 1. The presence of hexapeptides could be determined in hundreds of protein sequences. To illustrate, hexapeptide VLDVGG has been found in 2 human and 380 other protein sequences [34]. Radauer et al. [56] observed that experimentally determined allergens belong to a limited number of protein families. The above study relied solely on the Pfam database, therefore the reported findings cannot be directly compared with our results. The order of magnitude of the number of proteins determined in our study is consistent with that given by Radauer et al. [56].

The distribution of hexapeptides according to the number of proteins belonging to individual families and containing the query sequence is presented in Table 1. Most families contain single proteins with the searched fragments. Their presence in such sequences can be interpreted as random. This interpretation is consistent with the findings of Lavelle and Pearson [40].

Homologs of Gad c 1 do not contain query hexapeptides. They were not observed in other proteins containing domains such as parvalbumin (InterPro ID: IPR008080) or parvalbumin alpha (PANTHER ID: PTHR11653). Those hexapeptide sequences were also absent in proteins belonging to larger families such as EF-hand-like

(InterPro IPR011992); EF-hand 2 (InterPro IPR018249), Calcium-binding EF-hand (InterPro IPR002048) or the EF-hand superfamily (SUPERFAMILY SSF47473). Parvalbumin sequences contain only a limited number of common (i.e. occurring in more than one homologous protein) epitopes, as demonstrated in the BIOPEP database. Due to high sequence variability, parvalbumins are the recommended markers of individual fish species [6]. Kuehn et al. [38] recommended the use of panels of parvalbumins from different species in tests of fish allergens.

A single family does not contain more than 52 proteins (identified based on the presence of individual domains) including one of the query hexapeptides in amino acid sequence. Among families containing at least 10 proteins with DEDGFY fragment within domain, SH3-related domains (InterPro IPR001452, InterPro IPR011511 and CATH G3DSA:2.30.30.40) are involved in protein–protein interactions such as cellular signaling. Ribosome maturation factor-binding proteins are classified as a separate family in the PANTHER database (PANTHER PTHR14234 and its subfamilies). The RIM-binding protein domain overlaps with SH3-related domains cataloged in InterPro and CATH. The AMP-binding domain (InterPro IPR000873, PANTHER PTHR11968) occurs in enzymes such as luciferase (defined by domain: CATH G3DSA:2.30.38.10), acetyl CoA-synthetase (SUPERFAMILY SSF56801) and long-chain fatty acid CoA ligase (PANTHER PTHR11968:SF44). Those domains overlap in individual protein chains. Intradiol ring-cleavage dioxygenases (InterPro IPR000627 and IPR015889) are microbial enzymes involved in the degradation of aromatic rings. The IADEDK hexapeptide is found (in at least 10 proteins containing domain) in enzymes which catalyze reduction reactions, including sulfhydryl groups and enzymes containing the thioredoxin-like fold (InterPro IPR012336), as well as in related domains such as thioredoxin fold (InterPro IPR012335) and arsenate reductase-like protein (InterPro IPR006660). Those three domains overlap in individual proteins. Proteins with the WD40 repeat-like-containing domain (InterPro IPR011046) are found in eukaryotes, and they are involved in signal transduction and transcription regulation in cell cycle control and apoptosis. Repeated WD40 motifs act as sites for protein–protein interactions. Proteins containing WD40 repeats participate in the formation of protein complexes and interplay between other proteins. Families cataloged in the PANTHER database (PTHR12816 and PTHR12816:SF5) are defined based on the presence of domains with the WD40 repeat-like-containing domain in individual proteins. The PTHR12816:SF5 subfamily is considered as retinoblastoma-binding. Families containing the largest number of proteins with the KGILSN fragment participate in nitrogen fixation (Nitrogenase/oxido-reductase, component 1 – InterPro IPR000510, Nitrogenase molybdenum-iron protein beta chain – InterPro IPR005976, CATH G3DSA:3.40.50.1980 and the “Helical backbone” metal receptor superfamily – SUPERFAMILY SSF53807). Domains defining those families overlap with one another in individual protein chains. Catalysis of phosphorylation reaction (kinase activity) is yet another biological function attributed to proteins containing the above hexapeptide. The following families and domains are annotated as kinase or kinase-like: protein kinase-like domain (InterPro IPR011009), phosphorylase kinase domain 1 (CATH G3DSA:3.30.200.20), protein kinase, catalytic domain (InterPro IPR000719), pyruvate kinase (InterPro IPR001697), pyruvate kinase alpha/beta (InterPro IPR015794), pyruvate kinase C-terminal-like (InterPro IPR015795) and protein kinase, ATP binding site (IPR017441). Domain PTHR24420, cataloged as “not named” in the PANTHER database, covers the above families and domains. Another pair of overlapping domains consists of PANTHER PTHR16291 and PfamB PB008588. Those domains are not named and not associated with a specific function.

**Table 1**  
Distribution of Gad c 1 epitopic hexapeptides in the universal proteome.

| Hexapeptide                                  | Number of domains/families                                                                         |    |                                                                                                                                                                                                                     |    | Number of species                                                                    |    |
|----------------------------------------------|----------------------------------------------------------------------------------------------------|----|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----|--------------------------------------------------------------------------------------|----|
|                                              | Categories defined according to number of proteins with the hexapeptide sequence within the domain |    | Categories defined according to percentage of proteins within a family containing individual hexapeptides as part of a domain (calculated for families with at least two protein sequences containing hexapeptide). |    |                                                                                      |    |
| DEDFGY<br>236 proteins, 125 domains/families | 1                                                                                                  | 72 | <0.01%                                                                                                                                                                                                              | 8  | Animal and plant species relevant for the food industry and/or agriculture           | 9  |
|                                              | 2–9                                                                                                | 39 | 0.01–0.1%                                                                                                                                                                                                           | 10 | Useful and potentially useful microorganisms in the food industry and/or agriculture | 6  |
|                                              | At least 10                                                                                        | 14 | 0.1–1%                                                                                                                                                                                                              | 13 | Human symbionts and commensals                                                       | 9  |
|                                              |                                                                                                    |    | 1–10%                                                                                                                                                                                                               | 13 | Human pathogens and parasites                                                        | 25 |
|                                              |                                                                                                    |    | >10%                                                                                                                                                                                                                | 7  |                                                                                      |    |
| IADEDK<br>142 proteins, 115 domains/families | 1                                                                                                  | 72 | <0.01%                                                                                                                                                                                                              | 12 | Animal and plant species relevant for the food industry and/or agriculture           | 1  |
|                                              | 2–9                                                                                                | 37 | 0.01–0.1%                                                                                                                                                                                                           | 18 | Useful and potentially useful microorganisms in the food industry and/or agriculture | 4  |
|                                              | At least 10                                                                                        | 6  | 0.1–1%                                                                                                                                                                                                              | 11 | Human symbionts and commensals                                                       | 18 |
|                                              |                                                                                                    |    | 1–10%                                                                                                                                                                                                               | 2  | Human pathogens and parasites                                                        | 23 |
|                                              |                                                                                                    |    | >10%                                                                                                                                                                                                                | 1  |                                                                                      |    |
| KGILSN<br>161 proteins, 116 domains/families | 1                                                                                                  | 73 | <0.01%                                                                                                                                                                                                              | 6  | Animal and plant species relevant for the food industry and/or agriculture           | 8  |
|                                              | 2–9                                                                                                | 29 | 0.01–0.1%                                                                                                                                                                                                           | 14 | Useful and potentially useful microorganisms in the food industry and/or agriculture | 5  |
|                                              | At least 10                                                                                        | 14 | 0.1–1%                                                                                                                                                                                                              | 16 | Human symbionts and commensals                                                       | 15 |
|                                              |                                                                                                    |    | 1–10%                                                                                                                                                                                                               | 3  | Human pathogens and parasites                                                        | 26 |
|                                              |                                                                                                    |    | >10%                                                                                                                                                                                                                | 4  |                                                                                      |    |
| SNADIK<br>128 proteins, 126 domains/families | 1                                                                                                  | 80 | <0.01%                                                                                                                                                                                                              | 6  | Animal and plant species relevant for the food industry and/or agriculture           | 1  |
|                                              | 2–9                                                                                                | 46 | 0.01–0.1%                                                                                                                                                                                                           | 24 | Useful and potentially useful microorganisms in the food industry and/or agriculture | 5  |
|                                              | At least 10                                                                                        | 0  | 0.1–1%                                                                                                                                                                                                              | 14 | Human symbionts and commensals                                                       | 10 |
|                                              |                                                                                                    |    | 1–10%                                                                                                                                                                                                               | 1  | Human pathogens and parasites                                                        | 20 |
|                                              |                                                                                                    |    | >10%                                                                                                                                                                                                                | 1  |                                                                                      |    |

Several domains mentioned above are attributed to the same function. The number of possible biological functions of proteins containing one of four query fragments is, in fact, lower than the number of domains. Most families indicated here contain a small fraction of proteins with query hexapeptide fragments. This fact does not suggest a correlation between the presence of hexapeptides and biological functions of proteins.

The percentage ranges of members of individual families containing hexapeptides from Gad c 1 are presented in Table 1. Most families contain 0.01–0.1 and 0.1–1% of proteins with the above hexapeptides. Category of protein families with more than 10% proteins containing query fragments includes mostly rare domains annotated in PANTHER and PfamB databases. The maturation factor-binding protein family (PANTHER PTHR14234 and its subfamilies) is most broadly represented with regard to the DEDGFY fragment, and it is followed by the testican 3 subfamily (PANTHER PTHR12036:SF4). The testican family consists of proteoglycans which contribute to the social behavior of cells and organism development [1]. Testican 3 expression has been observed in T cells during leukemia [31]. The function of the PfamB PB015521 family has not been identified. Only one family (Retinoblastoma binding protein 5 – PANTHER PTHR12816:SF5) contains more than 10% proteins with the IADEDK fragment. Families with a relatively high share of proteins containing the KGILSN fragment are classified on the basis of rare domains described in PANTHER and PfamB databases. They are not attributed to specific functions. The SNADIK fragment is present predominantly in domains containing

the rare lark RNA-Binding protein registered in the PANTHER database (PTHR24011:SF174).

Selected categories of species which synthesize proteins containing query hexapeptides are listed in Table 1. The investigated fragments are also found in proteins (in particular enzymes) produced by soil and marine microorganisms. Some protein sequences have been found in metagenomes, and they are not attributed to particular species. Human pathogens and parasites form the most abundant category among these considered here. It should also be noted that none of the query peptides are cataloged in the AntigenDB database [3] of antigens and epitopes which are potentially useful for vaccine design.

Human proteins which contain hexapeptides originating from Gad c 1 are listed in Table 2. Ribosome maturation factor-binding proteins (PANTHER PTHR14234 and its subfamilies) from various species are among the most common precursors of the DEDGFY fragment. According to the PANTHER database, they are involved in neurological system processes, signal transduction and synaptic transmission. Proteins containing the KGILSN fragment are uncharacterized protein sequences encoded by the C17orf85 gene. They contain alpha-beta plait domains tentatively attributed to the nucleotide-binding function (InterPro IPR012677).

#### 4. Discussion

The number of proteins annotated in the UniProt database is growing rapidly [61], and future screenings are likely to reveal more

**Table 2**  
Human proteins containing fragments identical to the hexapeptides of Gad c 1.

| Protein name                            | Accession No. in UniProt database | Location       |
|-----------------------------------------|-----------------------------------|----------------|
| DEDFGY                                  |                                   |                |
| Isoform 3 of RIMS-binding protein 2     | O15034                            | 130–135        |
| Putative uncharacterized protein RIMBP2 | C9JWN3                            | 115–120        |
| RIMS-binding protein 3A                 | Q9UFD9                            | 778–783        |
| RIMS-binding protein 3B                 | A6NNM3                            | 778–783        |
| RIMS-binding protein 3C                 | A6NJZ7                            | 778–783        |
| Isoform 2 of RIMS-binding protein 2     | O15034-2                          | 115–120        |
| KGILSN                                  |                                   |                |
| Uncharacterized protein C17orf85        | Q53F19                            | 12–17; 292–297 |
| Uncharacterized protein C17orf85        | CQ085_HUMAN                       | 292–297        |

proteins containing the analyzed hexapeptide fragments. The number of domains seems to be stable, and according to recent research, they are unlikely to grow rapidly [20,41]. Adjustments in the attribution of proteins to different families could be expected. The number of proteins belonging to various families could be overestimated due to false positive results in computer-aided function predictions [58]. Improvements in function prediction algorithms will contribute to the accuracy of research studies.

The results of studies investigating the distribution of query fragments in different protein families do not reveal regularity. The ribosome maturation factor-binding protein family seems to be the only exception. Analyses of the investigated fragment have shown that this family is characterized by a relatively high number and high share of proteins demonstrating local identity with allergen Gad c 1.

In their efforts to classify proteins based on their presence in domains, some authors have relied solely on the Pfam database [22,56], whereas others showed a preference for the InterPro system [28]. InterPro is a somewhat redundant source as it lists overlapping domains that are linked to the same function in the same protein chain. However, studies that limit their search to a single domain database, such as Pfam, could overlook rare domains and families which could prove useful for achieving their research objective.

The interspecies similarity of proteins at oligopeptide level, even in the absence of homology, seems to be a common phenomenon [8,11,32,33,35,36,63]. The correlation between allergy and defense against pathogens has been reviewed by Black [5], Hedayati et al. [25], Lin et al. [44], Goldman and Huffnagle [21], Hofmann and Abraham [26] as well as Rindsjö and Scheynius [57]. The local identity between proteins from pathogens and allergens may have some practical consequences. Epitope removal is a recommended step in peptide or protein-based drug design [33,50]. Kanduc [33] recommends the use of epitopes with minimal similarity to human protein fragments as vaccines. The above solution aims to prevent autoimmune reactions caused by vaccines, and it may be extended. The possibility of similarity between vaccine and allergen epitopes should also be minimized. The use of epitope sequences as queries in database searches would support the development of a list of potential allergy risk factors, keeping in mind that identity at hexapeptide level is the recommended minimum criterion of possible cross-reactivity [4,23]. Data mining methods are also generally recommended in vaccine design [9].

Proteins containing hexapeptides identical to those in Gad c 1 are also synthesized by plants and animals which are common food sources, including rice (*Oryza sativa*) or pig (*Sus scrofa*). Proteins synthesized by microorganisms which have numerous applications

in food biotechnology also contain the analyzed hexapeptides. An example of the above is the *Aspergillus oryzae* fungus which is used in traditional Japanese fermentation techniques and is considered to be safe by the World Health Organization (WHO) [37]. Another example is the *Eubacterium eligens* microbial species that colonizes the human gastrointestinal system [46]. Some microorganisms may be listed as both human symbionts and commensals and human pathogens and parasites. Examples of the above include microflora species which are commonly found in the human body but which could also cause opportunistic infections, such as *Corynebacterium urealyticum* [59].

A list of pentapeptides from allergenic epitopes present in the human proteome was published by Kanduc [32]. The oligopeptides shared by human and microbial proteomes were described by Trost et al. [63]. In this study, we discussed the hexapeptides shared by a known allergen and the human proteome.

## 5. Conclusions

Hexapeptides from the allergen Gad c 1 are prevalent in the universal proteome. They are present in numerous non-homologous proteins with various functions, in particular enzymes of microbial origin (function annotated on the basis of presence of appropriate domain). Hexapeptides considered in Immune Epitope Database as B cell epitopes are found in plants and animals which are food sources, in microorganisms which are human symbionts and commensals and in pathogenic microorganisms. In the last case we would like to point out that there is necessary to avoid coverage between epitopes from pathogens and from allergens. This goal may be achieved for instance by Allergome database screening using epitope sequences as a queries. The Immune Epitope Database appears to be the most extensive resource concerning epitopes. The dissection of epitope sequences into short fragments [32,35] and use of such fragments as a query sequences seems to be the most effective strategy for screening this database using its search engine.

## Conflict of interest

Authors do not declare conflict of interests.

## Acknowledgement

This paper was financed under a research grant from the University of Warmia and Mazury in Olsztyn.

## Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at <http://dx.doi.org/10.1016/j.peptides.2012.08.011>.

## References

- [1] Alliel PM, Perin J-P, Jollès P, Bonnet FJ. Testican, a multidomain testicular proteoglycan resembling modulators of cell social behavior. *Eur J Biochem* 1993;214:347–50.
- [2] Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 1997;25:3389–402.
- [3] Ansari HR, Flower DR, Raghava GPS. AntigenDB: an immunoinformatics database of pathogen antigens. *Nucleic Acids Res* 2010;38:D847–53.
- [4] Bindslev-Jensen C, Sten E, Earl LK, Crevel RWR, Bindslev-Jensen U, Hansen TK, et al. 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* 2003;41:81–7.
- [5] Black PN. Does atopy protect against enteric infections. *Allergy* 2005;60:30–4.

- [6] Carrera M, Cañas B, Piñeiro C, Vázquez J, Gallardo JM. Identification of commercial hake and grenadier species by proteomic analysis of the parvalbumin fraction. *Proteomics* 2006;6:5278–87.
- [7] Cuff AL, Sillitoe I, Lewis T, Clegg AB, Rentzsch R, Furnham N, et al. Extending CATH: increasing coverage of the protein structure universe and linking structure with function. *Nucleic Acids Res* 2011;39:D420–6.
- [8] Darewicz M, Dziuba J, Minkiewicz P. Computational characterisation and identification of peptides for *in silico* detection of potentially celiac-toxic proteins. *Food Sci Technol Int* 2007;13:125–33.
- [9] Davies MN, Guan P, Blythe MJ, Salomon J, Toseland CP, Hattotuwegama C, et al. Using databases and data mining in vaccinology. *Expert Opin Drug Discov* 2007;2:19–35.
- [10] Dessailly BH, Redfern OC, Cuff A, Orengo CA. Exploiting structural classifications for function prediction: towards a domain grammar for protein function. *Curr Opin Struct Biol* 2009;19:349–56.
- [11] Dummer R, Mittelman A, Fanizzi FP, Lucchese G, Willers J, Kanduc D. Non-self-discrimination as a driving concept in the identification of an immunodominant HMW-MAA epitopic peptide sequence by autoantibodies from melanoma cancer patients. *Int J Cancer* 2004;111:720–6.
- [12] Elsayed S, Bennich H. The primary structure of allergen M from cod. *Scand J Immunol* 1975;4:203–8.
- [13] Elsayed S, Apold J, Aas K, Bennich H. The allergenic structure of allergen M from cod I. Tryptic peptides of fragment TM 1. *Int Arch Allergy Appl Immunol* 1976;52:59–63.
- [14] Elsayed S, Ragnarsson U, Apold J, Florvaag E, Vik H. Allergenic synthetic peptide corresponding to the second calcium-binding loop of cod allergen M. *Scand J Immunol* 1981;14:207–11.
- [15] Elsayed S, Apold J, Holen E, Vik H, Florvaag E, Dybendal T. The structural requirements of epitopes with IgE binding capacity demonstrated by three major allergens from fish, egg and tree pollen. *Scand J Clin Lab Invest* 1991;51:17–31.
- [16] Elsayed S, Apold J. Allergenic structure of Allergen M from cod II. Allergenicity of the limited tryptic hydrolysis peptides of fragment TM 2. *Int Arch Allergy Appl Immunol* 1977;54:171–5.
- [17] Elsayed S, Apold J. Immunochemical analysis of cod fish allergen M: locations of the immunoglobulin binding sites as demonstrated by the native and synthetic peptides. *Allergy* 1983;38:449–59.
- [18] Finn RD, Mistry J, Tate J, Coghill P, Heger A, Pollington JE, et al. The Pfam protein families database. *Nucleic Acids Res* 2010;38:D211–22.
- [19] García BE, Lizaso MT. Cross-reactivity syndromes in food allergy. *J Invest Allergol Clin Immunol* 2011;21:162–70.
- [20] Godzik A. Metagenomics and the protein universe. *Curr Opin Struct Biol* 2011;21:398–403.
- [21] Goldman DL, Huffnagle GB. Potential contribution of fungal infection and colonization to the development of allergy. *Med Mycol* 2009;47:445–56.
- [22] Gonzalez MW, Pearson WR. RefProtDom: a protein database with improved domain boundaries and homology relationships. *Bioinformatics* 2010;26:2361–2.
- [23] Goodman RE. Practical and predictive bioinformatic methods for the identification of potentially cross-reactive protein matches. *Mol Nutr Food Res* 2006;50:655–60.
- [24] Hattori M, Taylor TD. The human intestinal microbiome: a new frontier of human biology. *DNA Res* 2009;16:1–12.
- [25] Hedayati MT, Pasqualotto AC, Warn PA, Bowyer P, Denning DW. *Aspergillus flavus*: Human pathogen, allergen and mycotoxin producer. *Microbiology* 2007;153:1677–92.
- [26] Hofmann AM, Abraham SN. New roles for mast cells in modulating allergic reactions and immunity against pathogens. *Curr Opin Immunol* 2009;21:679–86.
- [27] Hunter S, Jones P, Mitchell A, Apweiler R, Attwood TK, Bateman A, et al. InterPro in 2011: new developments in the family and domain prediction database. *Nucleic Acids Res* 2012;40:D306–12.
- [28] Iwaniak A, Dziuba J. Analysis of domains in selected plant and animal food proteins – precursors of biologically active peptides. *Food Sci Technol Int* 2009;15:179–91.
- [29] Jain E, Bairoch A, Duvaud S, Phan I, Redaschi N, Suzek BE, et al. Infrastructure for the life sciences: design and implementation of the UniProt website. *BMC Bioinformatics* 2009;10:136.
- [30] Jedrychowski L, Wróblewska B, Szymkiewicz A. State of the art on food allergens – a review. *Pol J Food Nutr Sci* 2008;58:165–75.
- [31] Kamioka M, Imamura J, Komatsu N, Daibata M, Sugiura T. Testican 3 expression in adult T-cell leukemia. *Leuk Res* 2009;33:913–8.
- [32] Kanduc D. Correlating low-similarity peptide sequences and allergenic epitopes. *Curr Pharm Des* 2008;14:289–95.
- [33] Kanduc D. Epitopic peptides with low similarity to the host proteome: towards biological therapies without side effects. *Expert Opin Biol Ther* 2009;9:45–53.
- [34] Kanduc D. A confrontation between proteomes. *Self/Nonself* 2010;1:255–8.
- [35] Kanduc D, Stufano A, Lucchese G, Kusalik A. Massive peptide sharing between viral and human proteomes. *Peptides* 2008;29:1755–66.
- [36] Kleter GA, Peijnenburg AACM. 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* 2002;2:8.
- [37] Kobayashi T, Abe K, Asai K, Gomi K, Juvvadi PR, Kato M, et al. Genomics of *Aspergillus oryzae*. *Biosci Biotechnol Biochem* 2007;71:646–70.
- [38] Kuehn A, Hutt-Kempf E, Hilger C, Hentges F. Clinical monosensitivity to salmonid fish linked to specific IgE-epitopes on salmon and trout beta-parvalbumins. *Allergy* 2011;66:299–301.
- [39] Kusalik A, Trost B, Bickis M, Fasano C, Capone G, Kanduc D. Codon number shapes peptide redundancy in the universal proteome composition. *Peptides* 2009;30:1940–4.
- [40] Lavelle DT, Pearson WR. Globally, unrelated protein sequences appear random. *Bioinformatics* 2010;26:310–8.
- [41] Levitt M. Nature of the protein universe. *Proc Natl Acad Sci USA* 2009;106:11079–84.
- [42] Lima TA, Auchincloss AH, Coudert E, Keller G, Michoud K, Rivoire C, et al. HAMAP: a database of completely sequenced microbial proteome sets and manually curated microbial protein families in UniProtKB/Swiss-Prot. *Nucleic Acids Res* 2009;37:D471–8.
- [43] de Lima Morais DD, Fang H, Owen JL, Rackham OJ, Wilson D, Pethica R, et al. SUPERFAMILY 1.75 including a domain-centric gene ontology method. *Nucleic Acids Res* 2011;39:D427–34.
- [44] Lin Y-T, Wang C-T, Chiang B-L. Role of bacterial pathogens in atopic dermatitis. *Clin Rev Allergy Immunol* 2007;33:167–77.
- [45] Lopata AL, Lehrer SB. New insights into seafood allergy. *Curr Opin Allergy Clin Immunol* 2009;9:270–7.
- [46] Mahowald MA, Rey FE, Seedorf H, Turnbaugh PJ, Fulton RS, Wollam A, et al. Characterizing a model human gut microbiota composed of members of its two dominant bacterial phyla. *Proc Natl Acad Sci USA* 2009;106:5859–64.
- [47] 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.
- [48] Mari A, Rasi C, Palazzo P, Scala E. Allergen databases: current status and perspectives. *Curr Allergy Asthma Rep* 2009;9:376–83.
- [49] 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;61:153–8.
- [50] Nagata S, Pastan I. Removal of B cell epitopes as a practical approach for reducing the immunogenicity of foreign protein-based therapeutics. *Adv Drug Deliv Rev* 2009;61:977–85.
- [51] O'Neil C, Helbling AA, Lehrer SB. Allergic reactions to fish. *Clin Rev Allergy* 1993;11:183–200.
- [52] Pérez-Gordo M, Cuesta-Herranz J, Maroto AS, Cases B, Ibáñez MD, Vivanco F, et al. Identification of sole parvalbumin as a major allergen: study of cross-reactivity between parvalbumins in a Spanish fish-allergic population. *Clin Exp Allergy* 2011;41:750–8.
- [53] Petrey D, Honig B. Is protein classification necessary? Toward alternative approaches to function annotation. *Curr Opin Struct Biol* 2009;19:363–8.
- [54] Ponting CP, Schultz J, Copley RR, Andrade MA, Bork P. Evolution of domain families. *Adv Protein Chem* 2000;54:185–244.
- [55] Poulsen LK, Hansen TK, Nørgaard A, Vestergaard H, Skov PS, Bindslev-Jensen C. Allergens from fish and egg. *Allergy* 2001;56(Suppl. 67):39–42.
- [56] 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;121:847–52.
- [57] Rindsjö E, Scheynius A. Mechanisms of IgE-mediated allergy. *Exp Cell Res* 2010;316:1384–9.
- [58] Schnoes AM, Brown SD, Dodevski I, Babbitt PC. Annotation error in public databases: misannotation of molecular function in enzyme superfamilies. *PLoS Comput Biol* 2009;5:e1000605.
- [59] Seth-Smith H, Walker A. Opportunity knocks. *Nat Rev Microbiol* 2008;6:652–3.
- [60] Shevchenko A, Valcu C-M, Jungueira M. Tools for exploiting proteomes. *J Proteom* 2009;72:137–44.
- [61] The UniProt Consortium. Reorganizing the protein space at the Universal Protein Resource (UniProt). *Nucleic Acids Res* 2012;40:D71–5.
- [62] Thomas PD, Kejariwal A, Campbell MJ, Mi H, Diemer K, Guo N, et al. PANTHER: a browsable database of gene products organized by biological function, using curated protein family and superfamily classification. *Nucleic Acids Res* 2003;31:334–41.
- [63] Trost B, Lucchese G, Stufano A, Bickis M, Kusalik A, Kanduc D. No human protein is exempt from bacterial motifs, not even one. *Self/Nonself* 2010;1:328–34.
- [64] Van Do T, Elsayed S, Florvaag E, Hordvik I, Endresen C. Allergy to fish parvalbumins: studies on the cross-reactivity of allergens from 9 commonly consumed fish. *J Allergy Clin Immunol* 2005;116:1314–20.
- [65] Vita R, Zarebski L, Greenbaum JA, Emami H, Hoof I, Salimi N, et al. The Immune Epitope Database 2.0. *Nucleic Acids Res* 2010;38:D854–62.
- [66] Yeats C, Lees J, Carter P, Sillitoe I, Orengo C. The Gene3D Web Services: a platform for identifying, annotating and comparing structural domains in protein sequences. *Nucleic Acids Res* 2011;39:W546–50.

Table 1S: Selected species producing proteins which contain hexapeptide fragments from fish parvalbumins.

| Hexapeptide sequence | Animal and plant species relevant for the food industry and/or agriculture                                                                                                                                                                 | Useful and potentially useful microorganisms in the food industry and/or agriculture                                                                                                                  | Human symbionts and commensals (mainly oral and gastrointestinal microorganisms)                                                                                                                                                                                                      | Human pathogens and parasites                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                 |
|----------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| DEDGFY               | <i>Bos taurus</i> ; <i>Gadus morhua</i> subsp. <i>callarias</i> ; <i>Gallus gallus</i> ; <i>Laccaria bicolor</i> ; <i>Nicotiana tabacum</i> ; <i>Oryza sativa</i> ; <i>Ricinus communis</i> ; <i>Siniperca chuatsi</i> ; <i>Sus scrofa</i> | <i>Aspergillus oryzae</i> ; <i>Coprinopsis cinerea</i> ; <i>Corynebacterium efficiens</i> ; <i>Corynebacterium glutamicum</i> ; <i>Lactobacillus helveticus</i> ; <i>Streptomyces bingchenggensis</i> | <i>Collinsella stercoris</i> ; <i>Corynebacterium urealyticum</i> ; <i>Dermacoccus</i> sp.; <i>Eubacterium eligens</i> ; <i>Eubacterium hallii</i> ; <i>Eubacterium ventriosum</i> ; <i>Faecalibacterium prausnitzii</i> ; <i>Ruminococcus obeum</i> ; <i>Streptococcus australis</i> | <i>Acinetobacter haemolyticus</i> ; <i>Acinetobacter johnsonii</i> ; <i>Arthroderma gypseum</i> ; <i>Aspergillus flavus</i> ; <i>Aspergillus terreus</i> ; <i>Bacillus cereus</i> ; <i>Bacillus mycoides</i> ; <i>Bacillus weihenstephanensis</i> ; <i>Burkholderia ambifaria</i> ; <i>Burkholderia graminis</i> ; <i>Burkholderia</i> sp.; <i>Corynebacterium urealyticum</i> ; <i>Helicobacter cinaedi</i> ; <i>Leishmania donovani</i> ; <i>Leishmania infantum</i> ; <i>Leptospira interrogans</i> ; <i>Mycobacterium ulcerans</i> ; <i>Nectria haematococca</i> ; <i>Nocardia farcinica</i> ; <i>Schistosoma japonicum</i> ; <i>Schistosoma mansoni</i> ; <i>Segniliparus rotundus</i> ; |

|        |                                                                     |                                                                                                                                       |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                         |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                         |
|--------|---------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
|        |                                                                     |                                                                                                                                       |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                         | <i>Segniliparus rugosus</i> ; <i>Trichinella spiralis</i> ; <i>Vibrio vulnificus</i>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                    |
| IADEDK | <i>Gadus morhua</i><br>subsp. <i>callarias</i>                      | <i>Agrobacterium radiobacter</i> ; <i>Bacillus thuringiensis</i> ;<br><i>Bifidobacterium longum</i> ;<br><i>Pichia guilliermondii</i> | <i>Abiotrophia defectiva</i> ; <i>Bacteroides</i> sp.; <i>Bacteroides thetaiotaomicron</i> ;<br><i>Bifidobacterium longum</i> ;<br><i>Clostridium nexile</i> ; <i>Clostridium</i> sp. L2-50; <i>Coprococcus catus</i> ;<br><i>Eubacterium rectale</i> ; <i>Gemella haemolysans</i> ; <i>Gemella moribillum</i> ;<br><i>Holdemania filiformis</i> ; <i>Leptotrichia goodfellowii</i> ; <i>Pichia guilliermondii</i> ;<br><i>Roseburia intestinalis</i> ; <i>Roseburia inulinivorans</i> ; <i>Ruminococcus obeum</i> ;<br><i>Ruminococcus torques</i> ;<br><i>Streptococcus australis</i> | <i>Abiotrophia defectiva</i> ; <i>Acinetobacter lwoffii</i> ; <i>Aggregatibacter segnis</i> ;<br><i>Arcobacter butzleri</i> ; <i>Aspergillus fumigatus</i> ( <i>Neosartorya fumigata</i> );<br><i>Bacillus cereus</i> ; <i>Campylobacter upsaliensis</i> ; <i>Chaetomium globosum</i> ;<br><i>Clostridium butyricum</i> ; <i>Eubacterium saburreum</i> ; <i>Fusobacterium ulcerans</i> ;<br><i>Gemella haemolysans</i> ; <i>Gemella moribillum</i> ; <i>Leptotrichia goodfellowii</i> ;<br><i>Nectria haematococca</i> ; <i>Neosartorya fischeri</i> ; <i>Pichia guilliermondii</i> ;<br><i>Plasmodium vivax</i> ; <i>Prevotella bergensis</i> ; <i>Streptococcus australis</i> ;<br><i>Streptococcus bovis</i> ; <i>Streptococcus gallolyticus</i> ; <i>Streptococcus vestibularis</i> |
| KGILSN | <i>Bos taurus</i> ;<br><i>Cucumis melo</i> ;<br><i>Gadus morhua</i> | <i>Agrobacterium radiobacter</i> ;<br><i>Brevibacillus brevis</i> ;                                                                   | <i>Alistipes shahii</i> ; <i>Anaerotruncus colihominis</i> ; <i>Bacteroides fragilis</i> ;<br><i>Clostridium scindens</i> ; <i>Eubacterium</i>                                                                                                                                                                                                                                                                                                                                                                                                                                          | <i>Acinetobacter haemolyticus</i> ;<br><i>Acinetobacter junii</i> ; <i>Acinetobacter</i> sp.;<br><i>Alistipes shahii</i> ; <i>Anaerotruncus</i>                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                         |

|        |                                                                                                                                                           |                                                                                                                                                                          |                                                                                                                                                                                                                                                                                                                                                   |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                             |
|--------|-----------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
|        | subsp. <i>callarias</i> ;<br><i>Gallus gallus</i> ;<br><i>Oryza sativa</i> ;<br><i>Prunus cerasus</i> ;<br><i>Ricinus communis</i> ;<br><i>Sus scrofa</i> | <i>Debaryomyces hansenii</i> ;<br><i>Paenibacillus polymyxa</i> ;<br><i>Pichia angusta</i>                                                                               | <i>hallii</i> ; <i>Eubacterium ventriosum</i> ;<br><i>Lactobacillus coleohominis</i> ;<br><i>Methanosphaera stadtmanae</i> ;<br><i>Parvimonas micra</i> ; <i>Pasteurella dagmatis</i> ; <i>Peptoniphilus lacrimalis</i> ;<br><i>Peptoniphilus sp.</i> ; <i>Proteus mirabilis</i> ;<br><i>Proteus penneri</i> ; <i>Streptococcus parasanguinis</i> | <i>colihominis</i> ; <i>Bacillus cereus</i> ;<br><i>Bacteroides fragilis</i> ; <i>Borrelia garinii</i> ;<br><i>Campylobacter lari</i> ; <i>Chlamydia trachomatis</i> ; <i>Chryseobacterium gleum</i> ;<br><i>Clostridium difficile</i> ; <i>Enterobacter cloacae</i> ; <i>Ixodes scapularis</i> ; <i>Klebsiella pneumoniae</i> ; <i>Legionella pneumophila</i> ;<br><i>Loa loa</i> ; <i>Parvimonas micra</i> ;<br><i>Pasteurella dagmatis</i> ; <i>Pediculus humanus</i> ; <i>Peptoniphilus lacrimalis</i> ;<br><i>Plasmodium falciparum</i> ; <i>Plasmodium vivax</i> ; <i>Proteus mirabilis</i> ; <i>Proteus penneri</i> ; <i>Yersinia enterocolitica</i> |
| SNADIK | <i>Gadus morhua</i><br>subsp. <i>callarias</i>                                                                                                            | <i>Lactobacillus johnsonii</i> ;<br><i>Leuconostoc kimchii</i> ;<br><i>Metarhizium acridum</i> ;<br><i>Metarhizium anisopliae</i> ;<br><i>Streptococcus thermophilus</i> | <i>Bacteroides caccae</i> ; <i>Bacteroides ovatus</i> ; <i>Bacteroides sp.</i> ; <i>Candida albicans</i> ; <i>Enhydrobacter aerosaccus</i> ; <i>Lactobacillus johnsonii</i> ;<br><i>Lactobacillus ruminis</i> ;<br><i>Parabacteroides distasonis</i> ;<br><i>Parabacteroides sp.</i> ; <i>Ruminococcus obeum</i>                                  | <i>Acinetobacter calcoaceticus</i> ;<br><i>Acinetobacter sp.</i> ; <i>Arthroderma gypseum</i> ; <i>Borrelia recurrentis</i> ;<br><i>Campylobacter concisus</i> ;<br><i>Campylobacter rectus</i> ; <i>Candida albicans</i> ; <i>Clostridium botulinum</i> ;<br><i>Clostridium tetani</i> ; <i>Corynebacterium lipophiloflavum</i> ; <i>Pediculus humanus</i> ;                                                                                                                                                                                                                                                                                               |

---

*Schistosoma japonicum*; *Schistosoma*  
*mansoni*; *Shigella dysenteriae*;  
*Streptococcus salivarius*;  
*Streptococcus vestibularis*; *Treponema*  
*denticola*; *Trichinella spiralis*;  
*Trichomonas vaginalis*; *Trypanosoma*  
*cruzi*

---