POLISH JOURNAL OF FOOD AND NUTRITION SCIENCES

http://journal.pan.olsztyn.pl
e-mail: pjfns@pan.olsztyn.pl

Pol. J. Food Nutr. Sci.
2010, Vol. 60, No. 2, pp. 139-146

COMPUTER DATABASES IN CLASSIFICATION AND CHARACTERISTICS OF PROTEINS AS A SOURCE OF BIOACTIVE PEPTIDES

Anna Iwaniak, Joanna Frydrychewicz

Chair of Food Biochemistry, Faculty of Food Science, University of Warmia and Mazury in Olsztyn, Poland

Key words: protein classification, bioactive peptides, databases, BIOPEP database, evolutionary similarity, protein structure and function

Classification of proteins as precursors of bioactive peptides is presented in this work. To achieve this aim, the worldwide available computer databases such as BIOPEP, CATH, PDB, and SCOP were applied. The main qualitative criterion to classify the proteins was the integrated coefficient of biological activity of protein *(C)* defined as a square root of the sum of squares of *(A)* for different activities divided by the number of activities, where *(A)* denotes the frequency of occurrence of fragments with a given activity in a protein sequence and is described as the number of fragments with a given activity divided by the number of amino acid residues of a protein chain taken for an analysis.

Taking into consideration the coefficient *(C)* calculated for 126 animal and plant proteins, three families were distinguished. In the family containing proteins – the poorest source of bioactive fragments, were *e.g.* leguminlike chains of pumpkin, ginkgo biloba isolated from primary endosperm, vicia faba, and faba bean. Proteins being the best source of bioactive fragments (*e.g.* proteins derived from milk, bovine and chicken meat and wheat) were classified into the 1st family.

It was found out that such a family classification is not identical with protein classification according to the criteria proposed and applied in the other computer databases. However, some proteins contained similar bioactive fragments within the sequence chains as well as possessed similar functions or structural motifs (*e.g.* TIM barrel motif). It can be presumed about the evolutionary similarity of proteins as a source of bioactive peptides.

INTRODUCTION

Bioactive peptides are known as the inactive fragments within their protein precursors, which after the enzymatic action act with the appropriate receptors to regulate body functions. Such peptides often function as regulatory compounds, hormone-like substances and play an important (beneficial or not) physiological role as well as contribute to the content of functional foods [Wu *et al.,* 2006]. The peptides with biological activity regulate metabolism, affect body mass, adjust blood pressure, prevent oxidation processes *etc.* [Wang & Gonzalez de Mejia, 2006, Iwaniak & Minkiewicz, 2008]. Many endogenous peptides are produced during gastrointestinal digestion of proteins provided with food to the body [Wu *et al.,* 2006]. In most of the cases, food-derived peptides have from two to nine amino acid residues and according to Kitts & Weiler [2003] the number of amino acid units may be extended to twenty. Milk and dairy products are found so far as the best precursors of bioactive peptides [Kamiński *et al.,* 2007], but there is a plenty of them in the other sources like *e.g.*: egg, fish, meat, bacteria [Yoshikawa *et al.,* 2003; Dziuba *et al.,* 2009].

Food products are subject to changes over the passing years, and their considerable value is often a result of biological and chemical information obtained *via* bioinformatic tools. Bioinformatics provides the suitable knowledge about the molecular basis of human health and disease [Desiere *et al.,* 2001; Minkiewicz *et al.,* 2008].

Many laboratories apply computer techniques to evaluate food components including proteins. Such techniques are often used for modeling the physicochemical properties of proteins, structure prediction, homology search, function-structure relationship. The basis of the computer analysis of biomacromolecules are databases coupled with the specially-designed algorithms, *e.g.* BIOPEP: http://www.uwm.edu.pl/biochemia [Dziuba & Iwaniak, 2006] – a database suitable in the evaluation of protein as a source of bioactive peptides; InterPro – a database of structural motifs: http://www.ebi.ac.uk/interpro [Apweiler *et al.,* 2001], and CATH – a database of the hierarchic classification of protein domain structures: http://www.biochem.ucl.ac.uk./bsm/cath [Bray *et al.,* 2000]. Much of the value of these resources are the part of the interconnected databases with the cross-references which provide the basis platform for more advanced data integration strategies [Whitfield *et al.,* 2006]. There are also many QSAR (quantitative structure-activity relationship) techniques used to analyse the structure–activity connections of a protein or a peptide by the mathematical interpretation of amino acid descriptors like hydrophobicity and molecular bulkiness [Pripp & Ardö, 2007]. For instance, by using the QSAR method the prolyl oligopeptidase in blood serum was found to influence the level of hormones and neuropeptides which are implicated in Alzeheimer's disease [Pripp, 2006]. In the QSAR analysis of peptides, the value of IC_{50}, *i.e.* the concentration of bioactive fragment(s) corresponding to its half-inhibitory activity, is usually the measure of the biological

Author's address for correspondence: Dr. Anna Iwaniak, Chair of Food Biochemistry, Warmia and Mazury University in Olsztyn, ul. Pl. Cieszyński 1, 10-726 Olsztyn, Poland; tel.: (48 89) 523 37 22; e-mail: ami@uwm.edu.pl

activity of a peptide. Such values were obtained both under *in vitro* and *in vivo* conditions [Wu *et al.,* 2006].

In this study, we tried to classify proteins based on the similarity between values of the integrated coefficient of protein biological activity *(C)* and then to compare such a classification with the other classifications obtained by the use of selected worldwide accessible databases.

MATERIALS AND METHODS

Protein sequences

We analysed 126 protein sequences described and available in the BIOPEP database (http://www.uwm.edu.pl/biochemia), and they were present under their ID numbers from 1076 to 1201 [Dziuba & Iwaniak, 2006].

In order to group of sequences that share the similar characteristics of bioactive peptides the following evaluation criterion was applied:

1) the integrated coefficient of protein biological activity *(C):*

$$C = \sqrt{\frac{(A_1)^2 + (A_2)^2 + \ldots (A_n)^2}{n}}$$

where: $A_{1\ldots n}$ – the occurrence frequency of fragments with a given activity (see below), and n – the number of activities [Dziuba & Iwaniak, 2003].

The *(A)* parameter denotes the occurrence frequency of bioactive fragments with a given activity and is described by the equation:

$$A = a/N$$

where: a – the number of fragments with a given activity in a protein chain, and N- the number of amino acid residues of a protein [Dziuba & Iwaniak, 2006].

The above-mentioned discriminant can be automatically generated by a BIOPEP database user by clicking the function called "A, B, Y calculation" [Minkiewicz *et al*., 2008].

The values of *(A)* discriminant necessary to compute *(C)* coefficients were calculated for twenty three activities such as: opioid agonist and antagonist, regulating ion flows, dipeptidyl peptidase IV inhibitors, embryotoxic, immunostimulating, antithrombotic, antiamnestic, antihypertensive, inhibitors of ubiquitin-mediated proteolysis, immunomodulating, bacterial permease ligands, neuropeptides, antioxidative, inhibitors of diprotin A and B, metal binding, antibacterial, chemotactic, smooth muscle contracting, antinociceptive, celiac toxic, and stimulating gamma-interferon production.

Amongst the 126 analysed sequences, the highest values of *(A)* parameter were obtained for the antihypertensive activity of milk proteins and the lowest *(A)* values for faba bean proteins (data not shown).

Comparison of protein classification according to other databases

The following computer databases were applied to find the similarities between proteins – the source of bioactive peptides:

a) SCOP (http://nar.oupjournals.org/cgi/content/full/28/1/25) [Gough & Chothia, 2002],

b) CATH (http://www.biochem.ucl.ac.uk/bsm/cath) [Orengo *et al.,* 1997],

c) PDB (http://www.rcsb.org/pdb/) [Berman *et al.,* 2000].

The SCOP (Structural Classification of Proteins) database classifies proteins with known structures including all entries from PDB according to different levels of the hierarchy. These levels include: family (clear evolutionary similarity), superfamily (probably common evolutionary ancestor) and fold (the same major secondary structure) [Andreeva *et al.,* 2008].

The main criteria of CATH database protein classification are: **C**lass (common secondary structure), **A**rchitecture (overall shape of domain structure), **T**opology (overall shape of domain structure with the connectivity of the secondary structure with the domain core) and **H**omology (common ancestor) [Orengo *et al.,* 1997]. In turn, Protein Data Bank (PDB) contains information about experimentally-determined structures of biomacromolecules [Berman *et al.,* 2000].

RESULTS AND DISCUSSION

The basis to group the proteins in the classes (families) according to the integrated coefficient of biological activity *(C)* was the calculation of the discriminant *(A)*, *i.e.* the occurrence frequency of fragments with biological activity in a protein sequence. The above-mentioned parameter *(A)* was successfully applied in our previous studies [Dziuba *et al.,* 2003a]. The limitation of protein classification based only on *(A)* values is the fact that it can be performed only for one activity at a time. Such a classification of proteins as a source of hypotensive peptides using the values of *(A)* was made by Iwaniak *et al.* [2005]. Introduction of the other criteria of protein evaluation as a source of bioactive peptides that might include the experimental measure of peptide activity such as *e.g.* IC_{50} values, could bring some obstacles. One of them is the fact that such values are not available for all activities of peptides, which limits the comparisons between the proteins – precursors of bioactive peptides [Dziuba & Iwaniak, 2006]. Thus, we introduced the coefficient *(C)* as the mathematical description of protein capability to be a good or a bad precursor of numerous peptides with a variety of activities.

The values of parameter *(C)* given in a descending order are shown in Table 1 and the general composition of protein groups with similar activities is present in Table 2 (three families). The division of proteins into families was performed taking into consideration the minimum of the function: number of proteins = f*(C)* (data not shown). It was performed by MS EXCEL'03 and allowed to obtain the two minima. They were the borderlines of the families and corresponded to the values of *(C)* equal to 0.0981 (alpha/beta-wheat gliadin, ID-1177) and to 0.0614 (alpha/beta-wheat gliadin precursor, ID-1182).

The first family includes proteins which can be the best (richest) precursors of bioactive peptides, namely *e.g.* bovine caseins, bovine elastin and collagen. The second group contains plant and animal proteins such as wheat gliadins, alpha s_1- bovine caseins (variants A, B, D) and sorghum kafirins.

TABLE 1. The values of the integrated coefficient of biological activity *(C)* of proteins analysed.

Protein	(C)	Protein	(C)
1. bovine β–casein, gen. var. A$_3$, *ID- 1099**	0.3176	2. bovine β–casein, gen. var. C, *ID-1101*	0.3061
3. bovine β–casein, gen. var. A$_2$, *ID-1098*	0.3046	4. bovine β–casein, gen. var. E, *ID-1102*	0.3045
5. bovine β–casein, gen. var. F, *ID-1103*	0.3023	6. bovine β–casein, gen. var. A$_1$, *ID-1097*	0.2986
7. bovine β–casein, gen. var. B, *ID-1100*	0.2880	8. bovine elastin, *ID-1076*	0.2429
9. bovine α 1- collagen (III), *ID-1111*	0.2417	10. bovine α 1- collagen (I) [fragment], *ID-1112*	0.2408
11. chicken α 1-collagen, *ID-1113*	0.2271	12. wheat glutenin, *ID-1110*	0.1344
13. bovine elastin, *ID-1107*	0.1005	14. α/β-wheat gliadin, *ID-1177*	0.0981
15. bovine α S$_1$-casein, *ID-1088*	0.096	16. bovine elastin, *ID-1194*	0.096
17. bovine α S$_1$-casein, gen. var. D, *ID-1089*	0.0937	18. bovine α S$_1$-casein, gen. var. B, *ID-1087*	0.0937
19. bovine α S$_1$-casein, gen. var. A, *ID-1086*	0.0898	20. α/β-wheat gliadin [precursor], *ID-1186*	0.0897
21. lamb β- lactoglobulin, *ID-1105*	0.0886	22. rice prolamin [precursor], CLONE PPROL 7, *ID-1152*	0.0884
23. sorghum kafirin PSKR2 [precursor], *ID-1196*	0.0878	24. bovine α S$_2$-casein, gen. var. A, *ID-1090*	0.0814
25. rice prolamin [precursor] CLONE PPROL 14, *ID-1154*	0.0809	26. human κ-casein, *ID – 1120*	0.0761
27. bovine κ-casein, *ID-1117*	0.0757	28. caprine β- lactoglobulin, *ID-1104*	0.0751
29. blueberry monellin, chain A, *ID-1170*	0.0745	30. bovine β- lactoglobulin, *ID-1116*	0.072
31. α/β-wheat gliadin MM1, [precursor], *ID-1179*	0.0667	32. sorghum kafirin PSK8, [precursor], *ID-1197*	0.0658
33. α/β-wheat gliadin [precursor], *ID-1178*	0.0651	34. sorghum kafirin PGK1, [precursor]*, ID-1149*	0.0649
35. caprine κ-casein, *ID-1109*	0.0638	36. α/β-wheat gliadin [precursor], *ID-1180*	0.0616
37. α/β-wheat gliadin [precursor], *ID-1181*	0.0615	38. α/β-wheat gliadin [precursor], *ID-1182*	0.0614
39. α/β-wheat gliadin [precursor], *ID-1183*	0.0614	40. α/β-wheat gliadin [precursor], *ID-1184*	0.0599
41. barley γ-hordein, [precursor], *ID-1150*	0.0579	42. soybean 13KD globulin, *ID-1160*	0.0565
43. α/β-wheat gliadin [precursor]*ID-1185*	0.0538	44. α/β-wheat gliadin [precursor], *ID-1147*	0.0526
45. human α- lactalbumin, *ID-1077*	0.051	46. wheat γ-gliadin, class B-III, [precursor], *ID-1145*	0.0491
47. wheat γ-gliadin, [precursor], *ID-1187*	0.049	48 wheat γ-gliadin, [precursor], *ID-1188*	0.0481
49. broad bean narbonin, fragment 1-17, *ID-1191*	0.0461	50. leech eglin C, *ID-1201*	0.0455
51. bovine α- lactalbumin, *ID-1115*	0.0453	52. caprine α-lactalbumin, *ID-1079*	0.0447
53. rat α- lactalbumin, *ID-1084*	0.0444	54. wheat γ-gliadin, [precursor], *ID-1146*	0.0437
55. moth lyzozyme, *ID-1093*	0.0434	56. human myosin, light chain, *ID – 1122*	0.043
57. human lactoferrin, *ID – 1121*	0.0428	58. ω-wheat gliadin, *ID-1189*	0.0428
59. phycocyanin, *ID – 1126*	0.0427	60. cocoa seed storage protein, *ID-1114*	0.0427
61. wheat γ-gliadin class B-I, [precursor], *ID-1148*	0.0426	62. chicken connectin 1, *ID – 1118*	0.0424
63. lamb α- lactalbumin, *ID-1082*	0.0424	64. ginkgo biloba β-leguminlike chain, *ID-1143*	0.042
65. pumpkin β-leguminlike chain, *ID-1142*	0.042	66. guinea pig α- lactalbumin, *ID-1169*	0.0407
67. soybean 11S globulin [precursor], *ID-1161*	0.0405	68. garden pea β-leguminlike chain, *ID-1158*	0.0396
69. human lyzozyme, *ID-1091*	0.0392	70. soybean seed storage 11S globulin, *ID-1163*	0.0381
71. chicken troponin, *ID-1135*	0.0381	72. chicken troponin, *ID-1136*	0.038
73. N- terminal fragment of β-lupin, *ID-1192*	0.0377	74. chicken troponin, *ID-1137*	0.0372
75. horse α- lactalbumin, *ID-1078*	0.0359	76. chicken troponin, *ID-1138*	0.0346
77. bilin binding protein – BBP, *ID-1199*	0.0346	78. soybean 12S globulin, [precursor], *ID-1167*	0.0345
79. barley γ1-purothionin, *ID-1172*	0.0344	80. barley γ2-purothionin, *ID-1132*	0.0344
81. retinol binding protein – RBP, *ID-1198*	0.0343	82. soybean basic 7S subunit globulin [precursor], *ID-1162*	0.0338
83. barley γ-hordothionin, *ID-1175*	0.0333	84. bovine κ-casein, *ID-1106*	0.0333
85. rice prolamin [precursor], CLONE PPROL 17, *ID – 1153*	0.033	86. chicken myosin, fragment 1-930, *ID – 1123*	0.0322
87. epidermal retin acid binding protein-EBP, *ID-1200*	0.0321	88. soybean seed storage 12S globulin, *ID-1165*	0.0319
89. camel α-lactalbumin, *ID-1085*	0.0318	90. uppland cotton legumin A, *ID-1164*	0.0316
91. faba bean β-leguminlike chain, *ID-1159*	0.0313	92. oat β-leguminlike chain, *ID-1151*	0.0313
93. rice β-leguminlike chain, *ID-1139*	0.0313	94. chicken myoglobin, *ID – 1125*	0.0312

continued on the next page

TABLE 1. Continued.

Protein	(C)	Protein	(C)
95. oat 12S seed storage globulin [precursor], *ID-1166*	0.0309	96. blueberry monellin, B chain, *ID-1083*	0.0308
97. flavodoxin, *ID-1108*	0.0307	98. plastocyanin, *ID-1127*	0.0296
99. azurin, *ID-1096*	0.0283	100. octopus α-lactalbumin, *ID-1081*	0.0283
101. odorant binding protein – OBP, *ID-1193*	0.0278	102. chicken lyzozyme, *ID-1092*	0.0276
103. rabbit lysozyme *ID-1119*	0.0275	104. dog lysozyme, *ID-1094*	0.0223
105. pigeon lysozyme, *ID-1095*	0.0223	106. murine protein – MUP, *ID-1195*	0.0218
107. barley α1-purothionin [precursor], *ID-1171*	0.0189	108. chicken β-tropomyosin, *ID-1130*	0.0187
109. chicken α-tropomyosin, *ID-1128*	0.0187	110. chicken β-tropomyosin, *ID-1129*	0.0187
111. barley α-hordothionin [precursor], *ID-1176*	0.0172	112. chicken myosin fragment 931 – 1921, *ID- 1124*	0.0167
113. chicken troponin, *ID-1131*	0.0167	114. rice 10KD prolamin [precursor], *ID-1168*	0.0163
115. rabbit α-lactalbumin, *ID-1080*	0.0147	116. barley α-purothionin [precursor], *ID-1174*	0.0142
117. soybean β-leguminlike chain, *ID-1157*	0.014	118. sunflower β-leguminlike chain, *ID-1156*	0.014
119. common flax β-leguminlike chain, *ID-1144*	0.014	120. rapeseed β-leguminlike chain, *ID-1141*	0.014
121. mouseaer cress β-leguminlike chain, *ID-1140*	0.014	122. rice prolamin [precursor], CLONE PPROL 4A, *ID-1155*	0.0127
123. barley A-I purothionin, *ID-1173*	0.0123	124. porcine troponin, *ID-1134*	0.0109
125. chicken troponin C, *ID-1133*	0.0058	126. broad bean 2S narbonin, fragment 1-12, *ID-1190*	0.0103

*BIOPEP identification number

TABLE 2. Families of proteins based on *(C)* discriminant as the criterion of classification (according to the minimum of the function: $N^* = f(C)$).

Source	Protein
FAMILY I (ranges of *(C)* = 0,0981-0,317)	
Bovine *(Bos taurus)*	β-casein (gen. var. A$_2$, E, A$_3$, C, B, A$_1$, F), elastins, α1-collagen
Chicken *(Gallus gallus)*	α1-collagen
Wheat *(Triticum aestivum)*	glutenin, α/β-gliadins
FAMILY II (ranges of *(C)* = 0,0614-0,0981)	
Bovine *(Bos taurus)*	αS$_1$-casein (gen. var. A, B, D), αS$_2$-casein, elastins
Wheat *(Triticum aestivum)*	gliadin precursors (α/β- and γ-gliadins)
Sheep *(Ovis aries)*	β-lactoglobulin
Caprine *(Capra hircus)*	
Rice *(Oryza sativa)*	prolamin precursors
Sorghum *(Sorghum vulgare)*	kafirins
Human *(Homo sapiens)*	αS$_2$-casein, κ -casein, α-lactalbumin
Barley (*Hordeum vulgare*)	hordothionins
Soybean *(Glycine max)*	globulins
FAMILY III (ranges of *(C)* = 0,0058-0,0614)	
Wheat *(Triticum aestivum)*	γ-gliadin precursors, α/β-gliadins
Leech *(Hirudo medicinalis)*	eglin C
Broad bean *(Vicia faba)*	narbonins
Bovine *(Bos taurus)*	α-lactalbumin, odorant binding protein (OBP), retinol binding protein (RBP)
Caprine *(Capra hircus)*	α-lactalbumin
Horse *(Eqqus caballus)*	
Sheep *(Ovis aries)*	
Rat *(Rattus norvegicus)*	
Pigeon *(Columba livia)*	
Camel (*Camelus dromedarius*)	
Rabbit *(Oryctolagus cuniculus)*	
Octopus *(Octopus vulgaris)*	

continued on the next page

TABLE 2. Continued.

Source	Protein
Human *(Homo sapiens)* Chicken *(Gallus gallus)* Dog *(Canis familiaris)* Moth *(Bombyx mori)*	lysozyme
Chlorophyll	phycocyanin
Cocoa	storage protein
Chicken *(Gallus gallus)*	connectin
Gingko biloba (*Ginkgo biloba)* Pumpkin *(Cucurbita species)* Garden pea *(Pisum sativum)* Faba bean *(Vicia faba)* Oat *(Avena sativa)* Rice *(Oryza sativa)* Soybean *(Glycine max)* Sunflower *(Helianthus annuus)* Flax *(Linum usitatissimum)* Rapeseed *(Brassica napus)*	β-leguminlike chain
Mouseear cress *(Arabidopsis thaliana)*	β-leguminlike chain, epidermal binding protein (EBP)
White butterfly *(Pieris brassicae);*	bilin binding protein (BBP)
Rat *(Rattus norvegicus)*	male urinary protein (MUP)
Eucaryota	Plastocyanin
Bacteria	Favodoxin
Bacteria *(Alcaligenes faecalis)*	Azurin

*number of proteins

It is commonly known that milk proteins are the best-known source of peptides with biological activities [Kamiński *et al.,* 2007], but our results show also which proteins can be "comparable" to milk-derived sequences in terms of bioactive fragments content. It is consistent with the theory of Karelin *et al*. [1998] that proteins involved in a variety of functions in the system can also be precursors of biologically-active peptides.

The third family (the worst and poorest source of bioactive peptides) contains leguminlike chains of pumpkin, pea, rice and ginkgo biloba isolated from the primary endosperm [Häger *et al.,* 1995]. Such a division of proteins differs from their traditional classification, in which the main attention was paid to their structure or evolutionary similarity. Thus, it may explain the presence of *e.g.* gliadins in all distinguished groups.

Although the protein classification based on the coefficient *(C)* values does not include the experimental measures of biological activity of peptides encrypted in the protein sequences, it can still be suitable for protein evaluation. The better the source of bioactive peptides the higher the probability to release them from the precursor [Dziuba & Iwaniak, 2006], which may be important in the formulation of bioactive food products. The food based on protein-derived peptides becomes a subject of growing commercial interests on the health-promoting markets and gives a basis for the novel concept of "personalized nutrition" [Korhonen & Pihlanto, 2006].

The grouping of the proteins from the BIOPEP according to the criteria proposed in other databases, like SCOP, CATH and PDB, was possible only in the case of fourteen sequences gathered in our database. It was due to the fact that abovementioned databases possess only three dimensional structures of well-known proteins. Despite the limitation of the number of protein sequences to compare (fourteen out of 126 input sequences), it was still worth to probe if there are some similarities between the proteins we usually find as functionally distant.

All fourteen sequences were accessible in the PDB. Five of them: human α-lactalbumin, flavodoxin, chicken troponin, bilin binding protein, eglin C were available in the SCOP and CATH databases. The remaining nine of protein sequences could be classified only by CATH database (azurin, lactoferrin, phycocyanin) or by SCOP (caprine α-lactalbumin, chicken myosin, plastocyanin, narbonin, murine protein, vitamin A binding protein). The results of the SCOP and CATH classification are shown in Tables 3 and 4, respectively. Milk proteins such as human and caprine alpha-lactalbumin are in α + β class which classifies them to lysozymes. Meat proteins, such as myosin and troponin, belong to α class with the characteristic EF motif, *i.e.* calcium-binding motifs composed of two helixes (E and F) connected with a loop. Calcium is bound by a loop region. Many proteins with EF hand motifs are regulated by calcium, which enables classifying them to the calmodulinlike family [Branden & Tooze, 1999].

TABLE 3. SCOP protein classification.

Protein	Source	SCOP classification			
		Class (C)	Fold (F)	Superfamily (S)	Family (F)
α-Lactalbumin (1B90*, 1FKV)	Human/caprine	α+β	Lysozymelike	Lysozymelike	Lysozyme (type C)
Flavodoxin, 1FLN	Bacteria	α/β	Flavodoxinlike	Flavoproteinlike	Flavodoxinlike
Myosin (1BR1) and troponin (1TNW)	Chicken	All α	EF motif	EF motif	Calmodulinlike
Plastocyanin, (1JXG)	Eucaryotic proteins	All α	Cupredoxinlike	Cupredoxin	Azurinlike
Narbonin (1NAR)	Broad bean	α/β	TIM barrel	Glycosidase (trans)	Citin (type II)
MUP (1DF3)/ RBP (1AQB)	Rat/bovine	All α	Lipocalin	Lipocalin	Vitamin A binding protein
BBP (1BBP)	White butterfly	All β	Lipocalin	Lipocalin	Bilin binding protein
Eglin C (1ACB)	Leech	α+β	Serine protease inhibitor (CI-2 type)	Serine protease inhibitor (CI-2 type)	Serine protease inhibitor (CI-2 type)

*Protein Data Bank (PDB) ID.

TABLE 4. CATH protein classification.

Protein	CATH classification			
	Class (C)	Architecture (A)	Topology (T)	Homology (H)
Human α-lactalbumin, 1B90*	All α	Orthogonal spiral	Lysozyme	Hydrolase
Phycocyanin, 1F99			Globin	Phycocyanin
Chicken troponin, 1TNW			Recoverin	EF motif
Azurin, 1AIZ	All β	Sandwich	Immunoglobulinlike	Cupredoxin (copper binding protein)
Bilin binding protein (BBP), 1BBP		Barrel	Metaloproteinase inhibitor (subunit 1)	Vitamin A transporting
Eglin C, 1ACB			Trombin subunit H	Trypsinlike (serine proteases)
Flavodoxin, 1FLN	α/β	Three layer sandwich	Rossman fold	Electron transporting
Lactoferrin, 1BOL			D-maltodexin binding protein	Periplasmic (bindinglike)

*Protein Data Bank (PDB) ID.

Proteins involved in binding vitamin A and bilin are in the lipocalin family. Lipocalins are the extracellular proteins with chain length of 160-180 amino acid residues. They are involved in the binding of small, mostly hydrophobic molecules such as retinol, the formation of covalent or non-covalent complexes with other soluble macromolecules like bovine β-lactoglobulin (Blg), retinol binding protein (RBP), bilin binding protein (BBP), odorant binding protein (OBP), and epidermal retinol binding protein (EBP). Apart from this, lipocalins are transport proteins and possess a common barrel motif as well as a motif defined as "all α" [Flower *et al.,* 2000; Dziuba *et al.,* 2003b].

Another protein – narbonin – is characterised by the presence of the motif called α/β-barrel. This motif is common for about 10% of well-characterised enzymatic structures and is also known as the TIM barrel. It was discovered in triosephosphate isomerase [Farber, 1993], the enzyme participating in carbohydrates metabolism. Thirty enzymes with such a motif have been found so far. It confirms the hypothesis that secondary, and not the primary, structure of protein decides about the common evolutionary roots, and that the tertiary structure is the most conservative feature of protein and thus unaltered during the evolution [Kubicz, 1999]

Table 4 shows proteins classified by CATH database. The results obtained are consistent with those obtained from SCOP at the class level. Proteins from lipocalin superfamily, like RBP and BBP, have a barrel architecture and the topology of metalproteinase inhibitors. Another protein, *i.e.* eglin, has the same motifs at the architectural level.

The function of the proteins was described in Table 5 and analysedby using the Protein Data Bank (PDB) [Berman *et al.,* 2000]. According to PDB, troponins and human and caprine alpha-lactalbumins share the same function, *i.e.* they are calcium binding proteins. Phycocyanin and plastocyanin are involved in photosynthesis process and with the eglin C belong to hydrolases. Amongst the proteins analysed, some (lipocalins, eglin, narbonin) have the barrel motif. Researchers emphasize that proteins with this motif are very interesting from the scientific point of view. It is common knowledge that the members of the same family possessing a similar function and structure have to have a common ancestor. If the members of the same family serve a similar function but differ in the tertiary structure, it means that their evolution must have been convergent to form *e.g.* identical catalytic centre like in the case of serine proteases. In the case of possess-

TABLE 5. PDB protein classification.

Function	Protein
Calcium binding protein	Human α-lactalbumin, 1B90* Chicken troponin, 1TNW
Electron transporting	Flavodoxin, 1FLN Azurin, 1AIZ
Photosynthesis protein	Phycocyanin, 1F99 Plastocyanin, 1JXG
Transferase	Caprine α-lactalbumin, 1FKV
Hydrolase	Eglin C, 1ACB
Metal binding protein	Lactoferrin, 1BOL
Muscle protein	Chicken myosin, 1BR1
Seed storage protein	Broad bean narbonin, 1NAR
Transporting	Murine protein (MUP), 1DF3
Vitamin A transporting protein	Vit. A binding protein (RBP), 1AQB
Bilin binding protein	Bilin binding protein (BBP), 1BBP

*Protein Data Bank (PDB) ID.

ing the similar tertiary structure but different function (*e.g.* enzymes with alpha/beta-barrel domain), two evolutionary pathways are possible: (i) convergent evolution – the members of the family tend independently to adopt the solid type of ordered structure; and (ii) divergent evolution – members of the family have the common ancestor. In the second case, it has been assumed that the lack of homology of the sequence with the similar tertiary structure does not have to indicate no relationship. It may indicate the very ancient ancestry of the primal molecule, because the three-dimensional structure evolves slower than the primary one. Majority of researchers tend to accept the divergent evolutionary pathway of the family with the alpha/beta-barrel domain. It is the evidence of the common ancestry of proteins with different functions [Kubicz, 1999].

The classification obtained by calculating the integrated coefficient of biological activity of a protein is not consistent with the ones performed by the use of other databases. The explanation of this fact as well as the main obstacle is the lack of the 3D-structures for all the proteins analysed. We can confirm that some of the protein sequences possessed similar structural motifs like TIM barrel domain. It points to the evolutionary similarity of proteins being the source of bioactive peptides.

CONCLUSIONS

1. The introduction of protein evaluation criteria such as: the frequency of the occurrence of fragments with a given activity in a protein chain *(A)* and the integrated coefficient of biological activity of protein *(C)* can be helpful in the analysis of evolutionary relationships between proteins.

2. The higher the value of *(C)* parameter the richer the protein in the bioactive fragments, which gives three families of proteins. The best source of peptides with biological activity are milk proteins especially bovine beta-caseins, whilst the worse ones include porcine and chicken troponins.

3. There are no straight relations between the families of proteins based on discriminant (*C*) calculation and families classified according to structure similarity. It can be assumed that proteins with a similar activity profile can contain common structural motifs and, as a consequence, a common ancestor. To confirm this hypothesis recognition of all 3D-structures of proteins seems to be essential.

ACKNOWLEDGEMENTS

This work was supported by Project No. 528-0712-0809. The authors are grateful to Professor Jerzy Dziuba for the valuable contribution in the area of quantitative criteria used to evaluate food proteins as a source of bioactive peptides.

REFERENCES

1. Andreeva A., Howorth D., Chandonia J.-M., Brenner S.E., Hubbard T.J.P., Chothia C., Murzin A.G., Data growth and its impact on the SCOP database: new developments. Nucl. Acids Res., 2008, 36, D419-D425.
2. Apweiler R., Biswas M., Fleischmann W., Kanapin A., Karavidopoulou Y., Kersey P., Kriventseva E.V., Mittard V., Mulder N., Phan J., Zdobnov E., Proteome analysis database: on line application of InterPro and CluStr for the functional classification of proteins in whole genomes. Nucl. Acids Res., 2001, 29, 44–48.
3. Berman H.M., Westbrook J., Feng Z., Gilliland G., Bhat T.N., Weissig H., Shindyalov I.N., Bourne P.E., The Protein Data Bank. Nucl. Acids Res., 2000, 28, 235–242.
4. Branden C., Tooze J., Introduction to Protein Structure. 1999, Garland Publishing Inc., New York, p. 24.
5. Bray J.E., Todd A.E., Pearl F.M., Thornton J.M., Orengo C.A., The CATH dictionary of homologous superfamilies (DHS): a consensus approach for identifying distant structural homologies. Protein Eng., 2000, 13, 153–166.
6. Desiere F., German B., Watzke H., Pfeifer A., Saguy S., Bioinformatics and data knowledge: the new frontiers for nutrition and foods. Trends Food Sci. Technol., 2001, 12, 215–229.
7. Dziuba J., Iwaniak A., Database of protein and bioactive peptide sequences. 2006, *in:* Nutraceutical Proteins and Peptides in Health and Disease (eds. Mine Y., F. Shahidi). CRC Press, Taylor & Francis Group, Boca Raton, Florida, pp. 543–564.
8. Dziuba J., Iwaniak A., Food proteins as the source of bioactive peptides. 2003, Plenary lecture delivered at the 2nd National Congress of Biotechnology, 23–27 June, Łódź (in Polish).
9. Dziuba M., Dziuba B., Iwaniak A., Milk proteins as precursors of bioactive peptides. Acta Sci. Polon. Technol. Aliment., 2009, 8, 71–90.
10. Dziuba J., Iwaniak A., Minkiewicz P., Computer-aided characteristics of proteins as potential precursors of bioactive peptides. Polimery, 2003a, 48, 50–53.
11. Dziuba J., Iwaniak A., Niklewicz M., Minkiewicz P., Bovine β-lactoglobulin and other lipocalin as the source of bioactive peptides. Curr. Top. Protein Pept. Res., 2003b, 5, 101–104.

12. Farber G.K., An α/β–barrel full of evolutionary trouble. Curr. Opin. Struct. Biol., 1993, 3, 409–412.
13. Flower D.R., North A.C.T., Sansom C.E., The new lipocalin protein family: structural and sequence overview. Biochim. Biophys. Acta, 2000, 1482, 9–24.
14. Gough J., Chothia C. SUPERFAMILY: HMMS representing all proteins of known structure. SCOP sequence searches, alignments and genome assignments. Nucl. Acids Res., 2002, 30, 268–272.
15. Häger K.-P., Braun H., Czihal A., Mûller B., Bäumlein H., Evolution of seed storage protein genes: Legumin genes of *Ginkgo biloba*. J. Mol. Evol., 1995, 41, 457–466.
16. Iwaniak A., Dziuba J., Niklewicz M., The BIOPEP database – a tool for the *in silico* method of classification of food proteins as the source of peptides with antihypertensive activity. Acta Alim., 2005, 34, 417–425.
17. Iwaniak A., Minkiewicz P., Biologically active peptides derived from proteins. Pol. J. Food Nutr. Sci., 2008, 58, 289–294.
18. Kamiński S., Cieślińska A., Kostyra E., Polymorphism of bovine beta-casein and its potential effect on human health. J. Appl. Gen., 2007, 48, 189–198.
19. Karelin A.A., Philippova M.M., Karelina E.V., Strizhkov B.N., Grishina G.A., Nazimov I.V., Ivanov V.T., Peptides from bovine brain: structure and biological role. J. Pept. Sci., 1998, 4, 211.
20. Kitts D.D., Weiler K., Bioactive proteins and peptides from food sources. Applications of bioprocesses used in isolation and recovery. Curr. Pharm. Des., 2003, 9, 1309–1323.
21. Korhonen H., Pihlanto A., Bioactive peptides: production and functionality. Int. Dairy J., 2006, 16, 945–960.
22. Kubicz A., The secrets of molecular evolution. 1999, PWN, Warszawa, p. 15 (in Polish).
23. Minkiewicz P., Dziuba J., Darewicz M., Iwaniak A., Dziuba M., Nałęcz D., Food peptidomics. Food Technol. Biotechnol., 2008, 46, 1–10.
24. Orengo C.A., Michie A.D., Jones S., Jones D.T., Swindells M.B., Thornton J.M., Cath – a hierarchic classification of protein domain structures. Structure, 1997, 5, 1093–1108.
25. Pripp H.A., Quantitative structure-activity relationship of prolyl oligopeptidase inhibitory peptides derived from β-casein using simple amino acid descriptors. J. Agric. Food Chem., 2006, 54, 224–228.
26. Pripp A.H., Ardö Y., Modelling relationship between angiotensin-(I)-converting enzyme inhibition and the bitter taste of peptides. Food Chem., 2007, 102, 880–888.
27. Wang W., Gonzalez de Mejia E.G., A new frontier in soy bioactive peptides that may prevent age-related chronic diseases. Comp. Rev. Food Sci. Food Safety, 2006, 4, 63–78.
28. Whitfield E.J., Pruess M., Apweiler R., Bioinformatics database infrastructure for biotechnology research. J. Biotechnol., 2006, 124, 629–639.
29. Wu J., Aluko R.E., Nakai S., Structural requirements of angiotensin I-converting enzyme inhibitory peptides: quantitative structure-activity relationship study of di- and tripeptides. J. Agric. Food Chem., 2006, 54, 732–738.
30. Yoshikawa M., Takahashi M., Yang S., Delta opioid peptides derived from plant proteins. Cur. Pharm. Des., 2003, 9, 1325–1330.

Received February 2009. Revision received and accepted January 2010.