Uncovering conserved patterns in bioactive peptides in Metazoa

Uncovering conserved patterns in bioactive peptides in metazoa

Feng Liu [a], *Geert Baggerman* [b,*], *Liliane Schoofs* [b], *Geert Wets* [a]

[a] *Data Analysis & Modeling Group, Transportation Research Institute, Hasselt University, Building D, 3590 Diepenbeek, Belgium*
[b] *Laboratory of Developmental Physiology, Genomics and Proteomics, K.U. Leuven, Naamsestraat 59, B-3000 Leuven, Belgium*

ARTICLE INFO

Article history:
Received 7 July 2006
Received in revised form
21 August 2006
Accepted 21 August 2006

Keywords:
Neuropeptide
Toxin
Growth factor
Peptide precursor
Conserved pattern
PROSITE database
Pratt
Motif database
Peptide signature

ABSTRACT

Bioactive (neuro)peptides play critical roles in regulating most biological processes in animals. Peptides belonging to the same family are characterized by a typical sequence pattern that is conserved among the family's peptide members. Such a conserved pattern or motif usually corresponds to the functionally important part of the biologically active peptide. In this paper, all known bioactive (neuro)peptides annotated in Swiss-Prot and TrEMBL protein databases are collected, and the pattern searching program Pratt is used to search these unaligned peptide sequences for conserved patterns. The obtained patterns are then refined by combining the information on amino acids at important functional sites collected from the literature. All the identified patterns are further tested by scanning them against Swiss-Prot and TrEMBL protein databases. The diagnostic power of each pattern is validated by the fact that any annotated protein from Swiss-Prot and TrEMBL that contains one of the established patterns, is indeed a known (neuro)peptide precursor. We discovered 155 novel peptide patterns in addition to the 56 established ones in the PROSITE database. All the patterns cover 110 peptide families. Fifty-five of these families are not characterized by the PROSITE signatures, and 12 are also not identified by other existing motif databases, such as Pfam and SMART. Using the newly identified peptide signatures as a search tool, we predicted 95 hypothetical proteins as putative peptide precursors.

1. Introduction

Whole genome sequencing projects have made available immense sequence data at a pace that far supersedes their rate of annotation. As a result, out of 0.7 million protein sequences, which are currently available for all the completely sequenced metazoan genomes, nearly 18% could not be assigned any putative function. Although several tools/algorithms that contribute towards the putative functional assignments of the proteins are available, yet large numbers of proteins remain un-annotated. In most cases this is due to the low degrees of similarities with known proteins. Alternatively, the existing similarities can be confined to only very small part(s) of the entire protein. The latter is true especially for precursor proteins for bioactive peptides. Consequently, there is still a need for bioinformatic tools for predicting protein function to annotate the enormously large number of un-annotated protein sequences.

Bioactive peptides occur in the whole animal kingdom, from the least evolved phyla to the highest vertebrates. They play key roles as signalling molecules in many, if not all physiological processes, for instance as a peptidergic neurotransmitter or neurohormone, as a peptidergic toxin, or as a growth factor. Therefore, they are of considerable biological, medical and industrial importance [14,6]. Peptides are synthesized in the cell in the form of large preproproteins

* *Corresponding author.* Tel.: +32 16324260; fax: +32 16323902.
E-mail address: geert.baggerman@bio.kuleuven.be (G. Baggerman).
0196-9781/$ – see front matter © 2006 Published by Elsevier Inc.
doi:10.1016/j.peptides.2006.08.021

(precursors), which are then cleaved and modified to generate biologically functional peptides. Peptides (and their precursor proteins) that are structurally and functionally related have, since the elucidation of their structures, been classified in peptide families. A sequence motif or pattern common for a certain peptide family is often conserved throughout the whole animal kingdom. Usually the conserved peptide pattern resides in the functionally and structurally important part of the corresponding peptide precursor protein.

Conserved motif databases which cover a wide range of protein families exist in the form of patterns and profiles, such as PROSITE, Pfam and SMART [3,4,9,17]. The construction of these databases usually requires a good protein sequence alignment in order to produce accurate signatures. This works well when the sequences are easy to align. However, in some cases, the alignment is very difficult to obtain or evaluate, for example, when the conserved regions are very short or when they are repeated within the proteins. For peptide precursor proteins, in most cases, only a short conserved motif is responsible for the biological activity of their mature bioactive peptides and often only this short sequence motif, which can be five amino acids or less in length, is conserved [2]. This means that the bulk of a peptide precursor sequence is not very well preserved during evolution. The overall protein sequence identity, especially in distantly related homologues, may be too low for an accurate alignment.

In this paper, we have followed an alternative approach, taking unaligned sequences as a starting point. We then used a pattern search program to look for conserved patterns. We first collected all currently annotated peptide precursor proteins in Metazoa, through a search in Swiss-Prot and TrEMBL. The peptide sequences derived from these precursors are cleaved *in silico* and are classified into peptide family datasets. Then we use the pattern search program Pratt to search the peptide sequences in each peptide family dataset for conserved patterns. Such patterns consist of highly conserved positions that can be separated by fixed or variable spacing. The patterns are then refined by taking into account the information that is available in literature on the importance of amino acids contained within the biologically active site(s) of the peptide. The specificity of the generated patterns are further validated by scanning them against the protein databases Swiss-Prot and TrEMBL in order to ensure that proteins picked up by the pattern are either annotated as peptide precursor protein or have an unknown function.

2. Data collection

2.1. *Peptide precursors collection*

All proteins from metazoa that function as peptides or peptide precursor proteins, and that are further processed into smaller bioactive peptides are assembled into a peptide (precursor) database. A protein was considered as having the characteristics of a peptide or peptide precursor protein if it is annotated in Swiss-Prot or TrEMBL protein databases with one of the following keywords: hormone, antimicrobial, toxin. The hormone category includes subcategories of bombesin, bradykinin, cytokine, glucagon, growth factor, hormone, hypotensive agent, insulin, neuropeptide, neurotransmitter, opioid peptide, pyrokinin, tachykinin, thyroid hormone, vasoactive, vasoconstrictor and vasodilator (the definition of the keywords can be referred to in the two protein databases). The antimicrobial category consists of subcategories of antibiotic, antiviral defense, defensin and fungicide. The toxin includes naturally produced and secreted poisonous proteins that damage or kill other cells.

A protein is retained when its protein name or corresponding protein file contains one of the above-mentioned 24 peptide category or subcategory keywords in protein database UniProt (release 6.6) consisting of Swiss-Prot (release 48.6) and TrEMBL (release 31.6). However, when these proteins have a subcellular location as membrane protein (as indicated in the protein file) or if they are also characterized by keywords that do not refer to a peptide or peptide precursor protein, such as receptor, signal-anchor, transmembrane, binding protein, DNA binding, nuclear protein, transport, collagen, enzyme or words ending in 'ase' (excluding 'disease'), they are excluded, in order to avoid selection of proteins that are not peptide precursors. In total, 10,343 proteins are retained and they form the peptide (precursor) database.

Stand-alone PSI-BLAST is used to align all the peptide precursors with all the proteins in Swiss-Prot and TrEMBL databases except the ones that are already in the peptide precursor database. Based on the characteristics of peptide precursors that, in many cases, only a short motif within a peptide precursor is conserved and responsible for the function of the mature peptide [2], the score matrix PAM30 is used, and the word size is set to 2. This allows to find short but strong similarities. The proteins, which have similarities with the extracted peptide precursors with a significant BLAST score (e-value <0.001), are retained. The obtained list is then checked manually in terms of the protein's cellular location molecular function and biological process as stated by GO (gene ontology) terms or in literature. As a result, 1345 more proteins are added to the peptide precursor database and these proteins have as yet not been annotated by the peptide keywords in the protein databases.

2.2. *In silico extraction of peptides*

From each assembled peptide precursor, the bioactive peptide sequences are extracted in silico from the beginning and ending positions of the subsequences that are annotated as 'peptide' or 'chain' in 'feature' line in their corresponding protein files. The conserved basic cleavage sites flanking the peptides, which contribute to the cleavage of the peptides from their precursors, are also extracted along with the peptides. According to the characteristics of endoproteolytic peptide precursor processing [15], the rules applied to pick up the cleavage sites are proposed as follows. If the residues flanking the N or C-terminal of a peptide are the dibasic sites (G)KR, (G)RR, (G)KK or (G)RK, or the monobasic site (G)R or (G)K, the residues are extracted as cleavage sites. If the residues are a combination of a few consecutive basic amino acids K or R, the combination is extracted as the cleavage site.

Database entries in the peptide precursor database that only constitute the peptide sequence, i.e. in those cases where the protein precursor is unknown, are also retained. Small

proteins (less than 200 amino acids in length) from the peptide precursor database, which contain an N-terminal signal peptide and for which no mature peptides have as yet been annotated, presumably contain a single peptide and are therefore also extracted into the peptide database after in silico removal of the N-terminal signal peptide. According to statistics on all annotated bioactive peptide sequences, 97% are no longer than the 200 AA threshold value. The presence of a signal peptide is assumed when it is indicated on the protein file in the protein databases, in other cases the signal peptide prediction program signalP (http://www.cbs.dtu.dk/services/SignalP/) is used to predict the presence of a signal peptide.

In total, 12,697 peptide sequences are obtained and the sequences make up the peptide database, which we named 'Peptidedat'. The peptides are derived from 11,688 peptide precursors that originate from 1420 metazoan organisms. Most of the peptides are flanked by conserved basic cleavage sites.

2.3. *Peptide classification*

All the peptide or peptide precursor proteins in the peptide precursor database are classified into families according to their family classification information available in Uniprot that assign a protein to a particular family based on a significant match to an existing family motif, present in motif databases such as Prosite, Pfam and SMART or based on sequence similarities. Proteins that display sequence similarities with a significant score (e-value <0.001) obtained by BLAST, are clustered into the same family. A protein can also be assigned to belong to a particular peptide family based on its molecular function described in the literature. Once all the proteins in the peptide precursor database are classified into peptide families, the peptides within the peptide database 'Peptidedat' that are in silico cleaved from these precursors are automatically assigned into corresponding peptide family datasets.

In total, 110 datasets of peptide families are formed, each including at least 10 peptide sequences. Each of these families was scanned independently for conserved patterns.

3. Method

Different software tools available on the internet allow the user to search for patterns conserved in a set of unaligned protein sequences. Pratt (http://www.ebi.ac.uk/pratt/#) [10] is a flexible pattern search tool in the number of parameters that can be controlled by the user. It allows users to search for patterns of conserved positions with limited variable length spacing, which is important because even in well-conserved peptide regions, variable loop sizes can occur. Pratt is run on each of the 110 peptide family datasets. Based on the maximum of pattern length and pattern flexibility found in the existing peptide patterns in PROSITE, the parameter of the maximum pattern length (PL) is set to 52 amino acids, the maximum length of a wildcard is set to 15, the maximum number of flexible wildcards (FN) is 3, the maximum flexibility of a flexible wild card (FL) is 8 and the upper limit on the product of flexibilities for a pattern (FP) is 48, in order to allow more flexibility for the pattern to be searched for. The minimum percentage of sequences to match the pattern (C%) is set to 90, 80, 70, 60 and 50%, respectively. All other parameters are set at default.

For each Pratt run which starts with the parameter C% equal to 90%, the most significant pattern (the one with the highest fitness in the Pratt output list) is retained. The retained pattern is then refined by taking into account the important functional sites in the matched peptide sequences according to the corresponding literature. The amino acids occurring at these functional sites are added to the pattern if they are not included at the corresponding sites in the pattern.

The pattern is further corrected by scanning it against all proteins in Swiss-Prot and TrEMBL databases using the ScanProsite tool (http://www.expasy.org/tools/scanprosite/). Two possible cases occur: (1) If the pattern is not contained in any known non-neuropeptide protein, it is retained as a conserved peptide pattern. (2) Otherwise, if the pattern is matched by any annotated non-peptide proteins (further referred to as false positive protein hits), the correction processing is as follows. (2a) If the pattern does not include any wildcard region X, which is the site where any amino acid is accepted, the positions in all matching protein sequences where the pattern occurs are checked. If the pattern exclusively occurs at the N- or C-terminus of the matching peptide (precursor) proteins, or if all the matching peptide (precursor) proteins are exclusively small molecules, the pattern is retained with a constraint ('<' or '>') imposed at the N- or C-terminus of the pattern to limit the maximum distance between the conserved pattern region in the peptide or precursor proteins and the proteins' N- or C-terminus. If the pattern with such a restriction on its N- or C-terminus cannot distinguish these peptide or precursor proteins from non-peptide ones, the pattern is removed. (2b) Or, if the pattern has wildcard regions, all the matched sequences in both the false protein hits and the peptide precursor proteins are extracted and aligned. If the two groups of amino acids corresponding to a wildcard region X in the pattern have different physico-chemical properties between the false proteins and the peptide precursor proteins, the pattern is retained with a constraint being imposed at the X region that consists of the group of amino acids exclusively occurring in peptide precursor proteins. In the other case, when the two groups of amino acids share identical physico-chemical properties, the pattern is removed. The symbol sets from the amino acid class hierarchy (Smith and Smith, 1990): DE, KRH, NQ, ST, ILV, FWY, AG, C, M and P, which are classified based on the physiochemical nature of the side groups, are used.

If a conserved pattern cannot be obtained, the parameter C% is reduced by 10%, and Pratt is re-run against the peptide family dataset. As the percentage of sequences to match the pattern decreases, a pattern with more restrictions which is usually longer and contains more sites than the previously obtained pattern is shown up and processed by similar refinement and correction. Such a pattern may match less peptide sequences compared with the previously obtained one that is, however, unable to distinguish peptide precursors from non-neuropeptide proteins. The procedure is repeated until a pattern, which represents the majority of a group of

related peptide sequences, and which excludes any known non-peptide proteins, is discovered.

Once a conserved pattern is identified in the peptide family dataset, the program ps-scan (ftp://ftp.expasy.org/databases/prosite/tools/ps_scan/sources/) is run locally on the pattern against the dataset. The peptide sequences matching the pattern are extracted with the beginning and ending positions of the matched region. The matched sequence regions are removed from the corresponding original peptides. Each of the two remaining parts of the peptide sequences at their N- and C-terminus is retained to form an independent sequence if it is not less than four amino acids in length (given the assumption that the minimum length of the peptide pattern we search for is not less than this value). Thus, a reduced dataset is created including not only the peptides which are not covered by the identified pattern, but also the remaining sequences of the original peptides from which the sequence regions that already matched with a peptide pattern are removed. This methodology is based on the fact that a peptide precursor protein may contain several conserved regions, and that our established peptide database contains long peptide chains which may contain a few shorter, unrelated, bioactive peptides.

The reduced peptide family dataset is then scanned by Pratt to discover the next pattern. The search procedure is repeated again until the parameter C% is less than 50%. This means that the remaining dataset contains no more patterns representing the majority of the sequences.

It is important to note that instead of running Pratt only once to search for all significant patterns in the original peptide family dataset, the dataset is reduced by removing the sequences that match the previously discovered patterns, and is then scanned again. This is because in the Pratt output list, the front significant patterns often represent overlapping sites in peptide sequences. Therefore, a single scan would not be sufficiently sensitive to identify the next significant pattern located at a different region from the top significant one. By running it on a reduced dataset, Pratt is more efficient to discover the next novel significant pattern.

Fig. 1 represents the scheme of the described pattern searching procedure which is aimed to search short bioactive peptide sequences rather than their large precursor molecules, and to take into account not only the biologically functional sites of each individual peptide discussed in the literature, but also the general information which is extracted by the computational tool Pratt from all related peptides in a family.

4. Results

In total, we have identified 211 conserved peptide patterns, assembled in a peptide motif database, which we named 'peptidemotifdat'. These motifs represent 110 peptide families. All the 110 peptide families consist of 11,437 (98% of all proteins in the peptide precursor database) peptide or precursor proteins (12,400 peptides) in total, including 3716 cytokines and growth factors, 4974 hormones, 1265 antimicrobial proteins and 1482 toxins. Of all the proteins assembled in these 110 characterized peptide families, the identified patterns cover in total 10,715 (93.7%) peptide or precursor proteins (11,566 peptides) originating from species throughout the whole animal kingdom.

All the patterns are between 4 and 52 amino acids in length (excluding the constraints on protein length at the N-or C-terminus of the patterns), and 78 (37%) are no longer than 10 amino acids. Each of the patterns covers most of the peptide or peptide precursor proteins belonging to a family and the false positives are kept to zero because the peptide pattern is validated by the criterion that a known protein matching one of the patterns is a peptide or precursor protein in the corresponding family.

Tables 1 and 2 list all the currently identified patterns in PROSITE format, the name of the corresponding peptide family, the number of true positive proteins and false negative proteins, the hit number that reflects the repeated occurrence of the patterns within a peptide precursor protein and the novel peptide or peptide precursor proteins predicted by the patterns.

The largest families identified are TGF-beta and somatotropin hormone including 825 and 678 peptides or peptide precursors, respectively.

4.1. *Distribution of peptides*

Fig. 2 shows the length distribution of all peptides in the peptide database 'Peptidedat'. Almost all peptide sequences (98% according to the statistics on our peptide database) are shorter than 200 amino acids, and 1172 (9%) are between 3 and 10 amino acids. The minimum-length peptides are neuropeptide P83570 from *Sepia officinalis* which consists of only three amino acids GWamide [8] (The C-terminal amide assumes the presence of a Gly residue in the peptide precursor C-terminally of the peptide on which the alpha-amidating enzyme can act), human growth-modulating peptide P01157 (GHK) [16], and thyroliberin (QHP) found in many vertebrates [21]. The existence of numerous short bioactive peptides implies a very small conserved peptide motif can be a biologically important functional portion of the peptide precursor protein.

Fig. 3 shows the phylum distribution from which the assembled peptide sequences in our database originate. All assembled metazoan peptides are from various phyla including Chordata, Arthropoda, Annelida, Mollusca, Nematoda, Cnidaria, Echinodermata, Echiura, and Platyhelminthes. The majority of peptide sequence information is available within the group of the Chordata; 8877 peptide sequences have been identified, whereas in Echiura, only two peptides: urechistachykinin P40751 and P40752 from *Urechis unicinctus* are known to date.

4.2. *Comparison with the other motif databases*

The PROSITE database (http://ca.expasy.org/prosite) is a motif database of protein families and domains. It consists of biologically significant sites, patterns and profiles that help to reliably identify to which known protein family (if any) a new sequence belongs. Its 19.9 release contains 57 entries (patterns) describing 56 patterns for peptide families in metazoa (the omega-atracotoxin family has two patterns), and belonging to the following three categories including cytokines and growth factors, hormones and active peptides,

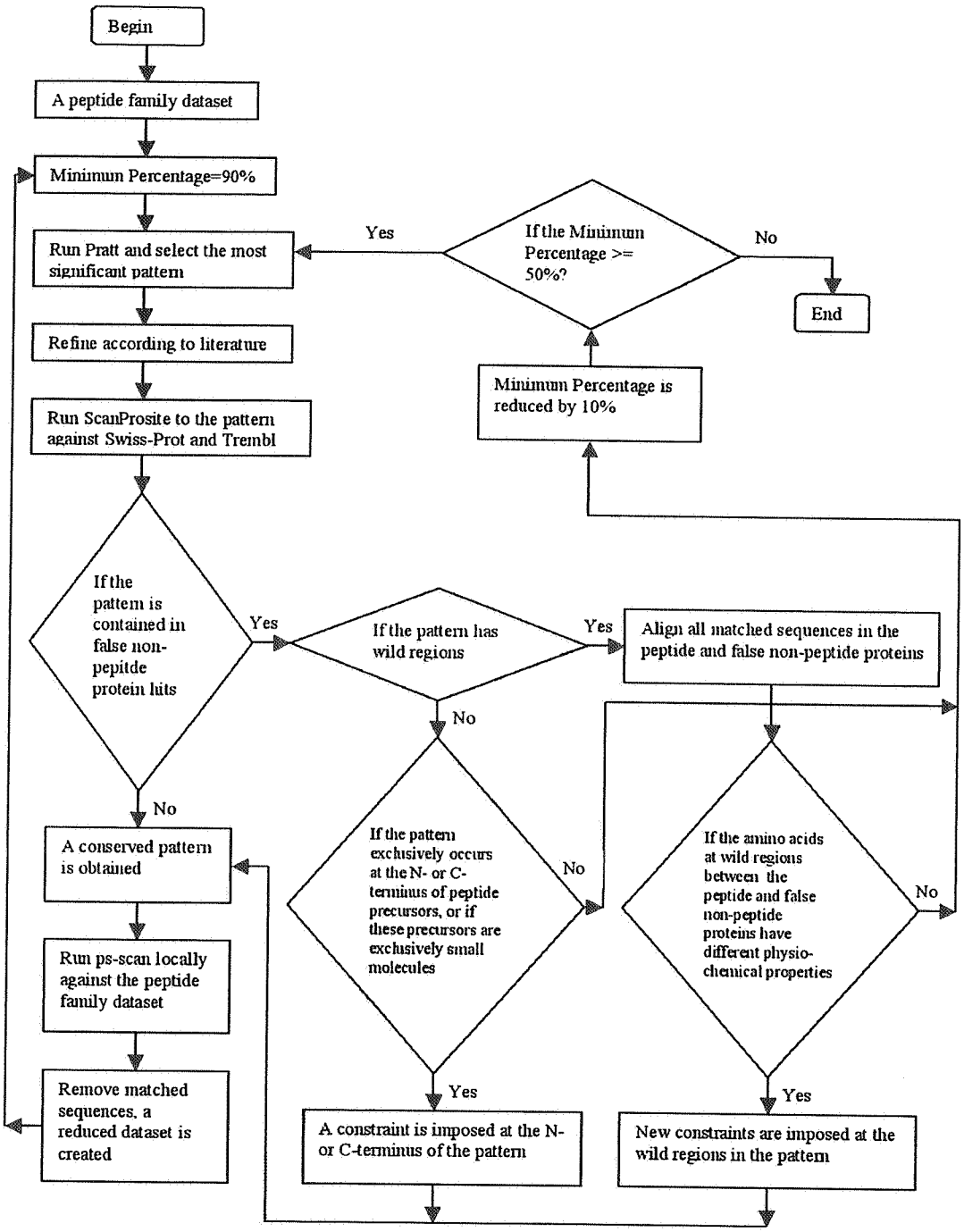

Fig. 1 – Program for identifying a peptide pattern.

and toxins. A 56 of the 211 peptide patterns discovered in the present study are similar to their PROSITE counterparts, but not identical (only five patterns were completely identical). The length and the conserved sites between the presently established peptide pattern and the PROSITE counterpart are similar, but in contrast to the PROSITE signatures, more amino acids are imposed at the conserved sites or wildcard regions in the currently established patterns. Compared with the PROSITE motifs, this will reduce both the number of false negatives and false positives, when using the novel pattern as a search tool. In addition, the novel peptide patterns are not only trained by running them against the Swiss-Prot protein database which is also used as the test dataset in PROSITE, but also against the TrEMBL database, in which many proteins are also annotated by keywords or literature.

Table 1 shows all the 56 peptide patterns that are found to be similar to the PROSITE signatures. Patterns (5) marked with 'identical' are completely identical to their PROSITE counter-

Table 1 – The conserved peptide patterns similar to PROSITE signatures

Name	Pattern	Tr. Po.	Hi.	Fa. Ne. Fr.	Fa. Ne. Pr.	New	Non-metazoa
Cytokines and growth factors							
(1) Granulins	(1) C-x-D-x(2)-H-C-C-[LIVM]-x(4)-C	42	241	2	0	Q616A1; Q7JKP2; Q9U362	1 (1)
(2) HBGF/FGF	(1) G-x-[LIVM]-[AGNP]-[STAGP]-[AGC]-[C]-x-[KRHNDE]-[WPC]-x-[STAGDENKRHQ](0,1)-[AGST](0,1)-[DENA]-C-[QP]-[FYLIVM]-[C]-[EQH]-x-[P]-[C]-[LIVM]-[DENKRHL]-[PLIVMDE]-[YHF]; (2) [GR]-[LIVM]-[LIVM]-[CWPDE]-[LIVM]-[PST]-[QLIVM]-x-[KRDEVIAGQFYNCS]-[STAGLMHQ]-[CP]-[AGDEN]-[FY]-[LIVM]-[AGSC]-[MLIV]-[NSTDEK]-[GAKRSTNDEQ]-[EDNKRHSTQA]-G (new); (3) G-S-[RHKQ]-[LIVM]-[CWPDE]-[LIVM]-[PST]-[QLIVM]-x-[KRDEVIAGQFYNCS]-[STAGLMHQ]-[CP]-[AGDEN]-[FY]-[LIVM]-[AGSC]-[MLIV]-[NSTDEK]-[GAKRSTNDEQ]-[EDNKRHSTQA]-G (new)	300	530	26	18		12 (30)
(3) PTN/MK heparin-binding	(1) S-[DE]-C-x-[DE]-W-x-W-x(2)-C-x-P-x-[SN]-x-D-C-G-[LIVMA]-G-x-R-E-G (identical); (2) C-[KR]-[YF]-x-[KRFY]-x(2)-W-[AGST]-x-C-[DENST] (new)	51	84	1	0		
(4) Nerve growth factor	(1) [GSRAED]-[CR]-[KRLIVM]-G-[LIVAT]-[DE]-[C]-x(2)-[YW]-[P]-S-x-[CR]; (2) [SAP]-[LIVA]-C-[DEY]-[SAG]-[WM]-[STDENC]-x-W-[VE]-[AGSTNI] (new)	321	471	11	1		3 (3)
(5) Platelet-derived growth factor (PDGF)	(1) P-[PSRAKQGL]-C-[LIVMFYAGST]-x(3)-[RQ]-C-[AGSTMLIVN]-G-S(0,1)-[CN]-C	158	158	17	6		16 (16)
(6) Small cytokines C-x-C	(1) C-x-C-[CFYW]-{CW}-x(3)-[P]-x(2)-[C]{8}-x(5,8)-C-x(2,3)-[EQMA]-[LIVMTE]-[LIVMF]-x(9,14)-C-[LIVMRK]-[DENH]	206	206	13	5	Q6DUZ6; Q6GLX8; Q4T8B9	71 (71)
(7) Small cytokines (intercrine/chemokine) C-C	(1) C-C-[LIVMFYSTQRKHDE]-[P](2)-{CDE}-[C]{7}-x(2,5)-[P]-[FYWAC]-[C]{2}-x(3,6)-C-{KM}-[C](1,3)-[SAG]-[LIVMTS]-[LIVMRTDE]-[FYLIVDE]-[C]{7,10}-C-[STAGVILM]	234	234	9	18	Q3ZBN3; Q32L58	9 (12)
(8) TGF-beta	(1) [WFYSTKRHL]-[LIVM]-[LIVMKRHF]-{CPNL}-P-[FY]-{PCW}-[FYILVA]-{C}-{QCWKRH}-{PA}-{PAGC}-C-{C}-{GE}-{C}-C	766	766	38	21		6 (7)
(9) Interferon alpha, beta and delta	(1) [FYH]-[FY]-{CP}-{GNRKCDSTI}-[LIVM]-{W}-{AGC}-{KRN}(0,1)-{FYLVIMN}-L-{PAG}-{C}-{PST}-{PFYW}-{FYHDEN}-x-{QY}-{CYQE}-{AT}-W; (2) L-{QKR}-x(0,4)-{GAEDVI}-{LVI}-{QHNDEFY}-{RQ}-{QH}-{LMIV}-{DENQVSTR}-x-L-{DENKRQ}-x-C-{LIVMKRQG} (new)	272	442	13	16		
(10) Granulocyte-macrophage colony-stimulating factor	(1) C-P-{LP}-T-{ST}-E-x-{QLIVMT}-C	25	25	1	0	Q4G094	
(11) Interleukin-1	(1) [LIVSTNDEFH]-{YESTMVIR}-{LFC}-{AGCFYL}-{SA}-{ASLV}-{CFY}-{CFYWH}-{PKRST}-{FYLC}-{WHLIVM}-{FYL}-{LI}-{SCA}-{TSVG}-x(6)-{PKRHCLIVMT}-x(0,2)-{LIVM}-{AGSTCVINDE}	128	128	16	8		
(12) Interleukin_2	(1) {ST}-E-{LF}-x(2)-L-x-C-L-x-{EDN}-E-L	74	74	14	0		

(13) Interleukin_4_13	(1) [LI]-x-E-[LIVM](2)-[Q](4)-x(0,1)-[LIVM]-[TL]-x(5,7)-C-x(2)-[LMIVST]-x-[IV]-x-[DNS]-[LIVMA]; (2) [KREV]-N-[STA]-[STED]-[DEAG]-{C}(3,4)-C-[RKT]-[AV]-[AGST]-[STKRHLVIQ]-x(9,15)-C (new)	73	119	3	1	
(14) Interleukin_6	(1) C-x(9)-C-[FYLIVM]-x(5)-G-L-x(2)-[FY]-x(3)-L	69	69	6	2	4 (7)
(15) Interleukin_7_9	(1) N-[DAT]-[LAPS]-[SCT]-F-L-K-[AGDE]-L-L	20	20	0	2	
(16) Interleukin_10	(1) [KQSN]-{C}(4)-C-[QYCH]-x(4)-[LIVM](2)-x-[FL]-[FYT]-[LMVRT]-x-[DERST]-[IV]-[LMF]	75	75	12	0	18 (21)
(17) LIF/OSM	(1) [PSTA]-x(4)-F-[NQ]-x-K-x(3)-[CG]-x-[LF]-L-x(2)-Y-[HK]	24	24	4	0	
(18) Osteopontin	(1) P-x(1,5)-[KQ]-x-[TA]-x(2)-[GA]-S-S-E-E-K	27	27	0	0	
Hormones						
(19) Adipokinetic	(1) [AGC]-Q-[LVI]-[NT]-[FY]-[ST]-[PASTKR]-[AGWSDEN]-W-[AGNDEST]; (2) <Q-[LVI]-[NT]-[FY]-[ST]-[PASTKR]-[AGWSDEN]-W-[AGNDEST>] (new)	45	45	0	0	Q5TTQ9
(20) Bombesin-like peptides	(1) [HLIVMQ]-W-A-[STIVRK]-G-[SH]-[LF]-M	42	42	1	0	
(21) Calcitonin/CGRP/IAPP	(1) [KR]-R-x(0,1)-C-[SAGDNT]-[STNG]-x(0,1)-[STAGVIL]-[TS]-C-[VMALI]-x(3)-[LYF]-x(3)-[LYFVI]; (2) <x(0,1)-C-[SAGDNT]-[STNG]-x(0,1)-[STAGVIL]-[TS]-C-[VMALI]-x(3)-[LYF]-x(3)-[LYFVI] (new)	83	84	7	0	
(22) Corticotropin-releasing factor	(1) [KR]-R-x(0,28)-[PQASLVIG]-[STPI]-[LIVM]-S-[LIVM]-x-[LIVMNAG]-[PST]-[LIVMFT]-x-[LIVM]-[LM]-[RN]-x(2)-[LIVMWF]; (2) <x(0,8)-[PQASLVIG]-[STPI]-[LIVM]-S-[LIVM]-x-[LIVMNAG]-[PST]-[LIVMFT]-x-[LIVM]-[LM]-[RN]-x(2)-[LIVMWF] (new); (3) T-R-[PQASLVIG]-[STPI]-[LIVM]-S-[LIVM]-x-[LIVMNAG]-[PST]-[LIVMFT]-x-[LIVM]-[LM]-[RN]-x(2)-[LIVMWF] (new)	64	64	4	5	Q4RWF4
(23) Arthropod CHH/MIH/GIH neurohormones	(1) [LIVM]-{C}-x(2)-C-[KR]-[FY]-[DENGRKHQ]-C-[FY]-{C}-[AGKRC]-{C}(2)-[FYILVM]-{C}-{CP}-C	135	135	5	0	Q23247
(24) Erythropoietin/thrombopoeitin	(1) P-x(4)-C-D-x-R-[LIVM](2)-x-[KRH]-x(14)-C	34	34	8	0	
(25) Granins	(1){DEF}-[DE]-[SN]-L-[SAN]-[AD]-[LIMVKR]-[DE]-[AGLSTQ]-E-L; (2) [LIVM]-x-[KHR]-C-[LIVM](2)-[ED]-[LIVM](2)-x(5)-[KRH]-[STP]-x(3)-[PST]-x(4)-C (new); (3) K-R-[STAG]-[NDEST]-[ED]-x(2)-[DE]-[DEGA]-[QKR]-Y-[AGST]-P-Q (new)	63	96	4	1	Q86T07; Q4RYY8; Q566G8
(26) Galanin	(1) G-W-[ST]-L-N-[ST]-[AG]-[AG]-[FY]-[LIVM]-[LIVM]-G-P; (2) <L-N-[ST]-[AG]-[AG]-[FY]-[LIVM]-[LIVM]-G-P (new)	31	31	0	1	
(27) Gastrin/cholecystokinin	(1) [FY]-x(0,2)-[GADN]-[AS](0,1)-[WH]-[MFLIV]-[DR]-F-G-[KR]-[RS]; (2) Y-x(0,2)-[GA]-[AS](0,1)-[WH]-[MFL]-[DR]-F> (new)	88	102	3	1	
(28) Glucagon/GIP/secretin/VIP	(1) [YH]-[STAIVGD]-[DENQ]-[AGF]-[LIVMSTE]-[FY]-[QLPAGDEKR]-[DENSTAK]-[DENSTA]-[LIVMFYG]-[RKSTDEN]-x(3)-[P]-[P]-x(2)-[AGSTLIVMQ]-[KREQL]-[KRDENQL]-[LVFYWG]-[LIVQ]	202	305	4	4	
(29) Glycoprotein hormones alpha chain	(1) C-x-G-C-C-[FY]-S-x-A-[FY]-P-T-P	109	109	1	3	
(30) Glycoprotein hormones beta chain	(1) C-{C}(2)-{CW}-{C}(7,9)-C-[STAGMLIVED]-G-[HFYLRS]-C-{C}-[STA]; (2) <x(0,8)-C-[STAGMDEVLI]-G-[HFYL]-C-{CKRH}-[ST] (new); (3) <x-{CW}-{C}(7,9)-C-[STAGDEVLIM]-G-[HFYL]-C-{C}-[ST] (new)	341	341	5	8	

Table 1 (*Continued*)

Name	Pattern	Tr. Po.	Hi.	Fa. Ne. Fr.	Fa. Ne. Pr.	New	Non-metazoa
(31) Gonadotropin-releasing hormones	(1) Q-[HY]-[FYW]-S-x(4)-P-G-G-[KR]-R; (2) Q-[HY]-[FYW]-S-x(4)-P-G> (new)	178	188	4	0		
(32) Insulin	(1) [C](2)-[IVLMPSTAFYR]-[CNE]-x-[C]-C-C-[CPM]-[P]-[CHW]-C-[STDNEKIGQ]-[C](2)-[CPAG]-[LIVMFSQ]-[CD]-[CPW]-[CHDEP]-C; (2) <x(0,205)-C-G-[FYILVMQW]-[CWPSTLIVM]-[LIVFY]-[VILMASTPH]-[AGHCFYPQW]-[CPQSW]-[LIVMRKHQWF]-[CNP]-[WCQP]-[LVIMATC]-C-[LM]-x(0,204)> (new)	507	877	31	21	Q32L79; Q621L6; Q61VN2; Q61GN7; Q4T1R8	
(33) Natriuretic peptides	(1) C-P-G-x(3)-[DEA]-[RH]-I-x(3)-[ST]-x(2)-G-C	155	155	5	5		
(34) Neurohypophysial hormones	(1) C-[LIFY]-[LIFYV]-x-N-C-P-x-G; (2) C-x(2,6)-[CW]-G-x(4,6)-C-[FYAGLIVM]-x(3)-[LIVFY]-C-C (new)	112	259	4	0		
(35) Neuromedin U and S	(1) [FY]-[LIVMF]-[FY]-R-P-R-N-G-[KR]; (2) [FY]-[LIVMF]-[FY]-R-P-R-N> (new)	24	24	3	0		
(36) Pancreatic	(1) [FY]-x(2)-[LIVM]-[LIVM]-x(2)-[YK]-x(3)-[LIVMFYRHK]-x-R-[PQVH]-R-[YF]-[GD]-[KR]-[RS]; (2) [FY]-x(3)-[LIVM]-x(2)-[YK]-x(3)-[LIVMFYRHK]-x-R-[PQVH]-R-[YF]-x(0,1)> (new)	118	118	4	3		
(37) Parathyroid hormone	(1) [KR]-R-x-[VI]-[STAGFYN]-[EH]-x-Q-x(2)-H-[DEN]-x-[GR]	54	54	2	1		
(38) Pyrokinins	(1) [AGHNQDEST]-[FYST]-[PQVIWFYED]-[FY]-[AGST]-P-R-[LI]-G-[KR]-R; (2) [AGHNQDEST]-[FYST]-[PQVIWFYED]-[FY]-[AGST]-P-R-[LI]> (new)	72	89	1	3	Q7PTL2; Q5TV14	
(39) Somatotropin	(1) C-[KRAG]-[STNRAC]-x(2)-[LIVMFYSRNW]-x-[LIVMSTAGY]-P-x(2)-[FYW]-x(2)-[TALIVMSHN]-x(7)-[LIVMFYP]-x(2)-[QHKR]-[KRHP]-[NW]-x-[LIVMFYR]-[LIVMSTC]-x-[STACVLMIG]-W; (2) C-[LIVMFG]-x-[KHRSNDEQVI]-[DEN]-[CNDEPQ]-[AGLMVI]-[KRMT]-[DENKRHPQ]-x-[STNALIVMF]-[FYLIVMKS]-[LIMVT]-x-[NDEKRH]-[LIVMATE]-[KRNEQTA]-C (new); (3) [ED]-K-L-L-[DE]-R-[VIA]-[IV]-x-H-[AT]-E-L (new); (4) C-F-[KRH](2)-[DEN]-[LIVMAG]-[HKR](2)-[LIVM]-[DEQ]-[ST]-[FYLIVM]-x(0,1)> (new)	633	1093	38	7		
(40) Tachykinin	(1) [AGSTQKRFY]-[SF]-[IVFYTHQ]-G-[LVIM]-M-G-[KR]-[RS]; (2) [AGSTQKRFY]-F-[IVLMFYSHQ]-G-[LVIMS]-R-G-K-R (new); (3) <x(0,9)-F-[IVLMFYTHQ]-G-[LVIMSTAG]-[RM]> (new)	104	124	3	3		
(41) (same)Urotensin II	(1) C-F-W-K-Y-C (identical)	30	30	0	1		
(42) Endothelin	(1) C-[C]-C-[C](4)-D-[C](2)-C-[C](2)-[FY]-C	50	104	2	0		
(43) Agouti	(1) C-[C](6)-C-[C](6)-C-C-[C](2)-C-[C](2)-C-[C]-C-[C](5,6)-C-[C]-C-[C](6,9)-C; (2) C-[C](6)-C-[C](6)-C-C-[C](2)-C-[C](2)-C-[C]-C-[C](5,6)-C-[C]-C-[C](0,8)> (new); (3) C-[C](6)-C-[C](6)-C-C-[C](2)-C-[C](2)-C> (new); (4) C-[C](6)-C-[C](6)-C-C-[C](2)-C-[C](2)-C-[C]-C-[C](5,6)-C(0,1)> (new)	37	37	7	0		
Antimicrobial							
(44) Cecropin	(1) W-[KDN]-[QNDEGAKRW]-[FYGA]-K-[KRE]-[LIVM]-E-[RKHAGN]-x-[AGVI]; (2) [GS]-[WRKHG]-[LIVMST]-[KRST]-K-[QNDEGAKRW]-[FYGA]-K-[KRED]-[LIVM]-E-[RKHAGN]-x-[AGVI] (new)	96	96	0	3	Q5TWE5	

(45) Mammalian defensins	(1) C-[C]-C-[C]{3,5}-C-[C]{6}-[CP]-[GARKSTW]-x-[SC]-[C]{6,10}-C-C; (2) C-[PR]-x-C-x{2,5}-C-x{2}-C-[PQ]-x-C-[PQ]-x-C (new)	119	145	2	3		
(46) Arthropod defensins	(1) [CG]-x{0,1}-[C]-[CQ]-[HNSEDRY]-C-x{3}-[C]{0,1}-[GR]-[A]-x-[GRQAY]-[GAL]-x-C-[FY]-x{3,4}-C-[C]-C; (2) [CG]-x{0,1}-[C]{2}-[HNSEDRY]-C-x{3}-[C]{0,1}-[GR]-[A]-x-[GRQAY]-[GAL]-x-C-[FY]-x{6}-C-[C]-C (new)	103	105	2	5	Q6XD83	3 (5)
(47) Cathelicidins	(1) Y-[LIVM]-[EDQN]-[AVI]-[LMVI]-[HKRG]-[RKHQ]-A-[LIVMA]-[DQGEN]-x-[LIVMFY]-N-[DEQ]	58	58	0	0		
Toxin							
(48) Snake toxins	(1) C-[CKRPL]-x{0,2}-C-[PRTFG]- [C]{5}-x{0,6}-C-C-[P]-x-[PDEN]-x-C-[NDEY]	352	352	4	16		
(49) Myotoxins	(1) K-x-C-H-x-K-x{2}-H-C-x{2}-K-x{3}-C-x{8}-K-x{2}-C	15	15	0	0		
(50) Scorpion short toxin 1	(1) C-[C]{4,5}-C-[PC]-[CQ]-[C]-C-x{3}-[C]-[CFWA]-x{1,4}-[GASEDN]-[KRAVISNDE]-C-[VIMQTDK]-[NG]-x{1,2}-[P]-C-[HKRDENVI]-C	77	77	4	2		
(51) Alpha-conotoxin	(1) <x{0,35}-[C]{15}-C-C-[SHYNDE]-[C]{2,3}-C-[C]{3,7}-C-[C]{0,12}>; (2) <[C]{0,14}-C-C-[SHYNDE]-[C]{2,3}-C-[C]{3,7}-C-[G>]> (new)	34	34	0	1		
(52) I-superfamily conotoxin	(1) C-[C]{6}-C-[C]{5}-C-C-[C]{1,3}-C-C-[C]{2,4}-C-[C]{9,10}-C (identical)	37	37	0	0		
(53) Mu-agatoxin and spider toxin SFI	(1) C-[C]{2}-[DEKR]-[C]{3}-C-[C]{4,7}-C-C-[C]{2,4}-C-[C]-C-[C]{4,15}-C-[C]-C-x{0,10}>	36	36	1	1		
(54) Omega-atracotoxin (ACTX) families	(1)C-[IT]-P-S-G-Q-P-C (identical); (2)C-C-[GE]-[ML]-T-P-x-C (identical)	13	13	0	0		
(55) Ergtoxin	(1) C-[C]{5}-C-x{8}-C-[C]{2}-C-C-x{9}-C-x{4}-C-[C]-C	25	25	0	0		

Tr. Po.: the number of true positive peptide or precursor proteins; Hi.: the number of matches to the pattern, and there may be more than a hit to the pattern within a protein; Fa. Ne. Fr.: the number of false negative protein fragments; Fa. Ne. Pr.: the number of false negative peptide precursor proteins; New: the novel putative peptide precursors belonging to the family predicted by the corresponding pattern; non-metazoa: the number of proteins in non-metazoa which match the peptide patterns. The number of all known non-metazoan proteins, which have similar molecular function to the metazoan proteins in the corresponding peptide family, is also listed in parentheses.

Table 2 – The novel conserved peptide patterns

Name	Pattern	Tr. Po.	Hi.	Fa. Ne. Fr.	Fa. Ne. Pr.	New	Non-metazoa
Cytokines and growth factors							
(1) Interferon gamma	(1) [RHSG]-[KRQ]-A-[AGFYLIVM]-x-[DE]-[LIVFY]-[QPAG]-x-[VI]-[VMLIY]-[LVIM]-x(1,4)-L-[STAGPKRLIVM]-[Q]-x(1,9)-[AGKR]-[KR]-R; (2) [RHSG]-[KRQ]-A-[AGFYLIVM]-x-[DE]-[LIVFY]-[QPAG]-x-[VI]-[VMLIY]-[LVIM]-x(1,4)-L-S-P-x(1,7)>	91	91	14	3		
(2) Interleukin_3	(1) [CVLIM]-[LIVM]-P-x-[AGPST]-x(2)-[STAGDENRKH]-x(12,14)-[DE]-F-[RKQ]-[NDEAGQST]-K-L	20	20	0	0		
(3) Interleukin_5	(1) [HDE]-x(2)-C-x(3)-[IVLM]-F-x-G-[LIVMST]-x(2)-L-x-[NST]	23	23	1	0		
(4) Interleukin_12 alpha	(1) [KRHE]-[LM]-C-x(2)-[LM]-[KRHQ]-[AG]-x(3)-R-x(2)-T-x(2)-[KR]-x(3)-Y-[LMIV]	34	34	7	0		
(5) Interleukin_15	(1)C-{C}(4)-[LM]-{C}-C-[FY]-[LIVFYQ]-x-[DE]-[LIVM]-x(2)-[LIVM]-x(2)-[ED]	44	44	1	0		
(6) Interleukin_17	(1) [RLM]-[QKR]-[PS]-{P}-x-[LIVMFY]-[RKH]-{CP}-[AS]-x-C-x-[CHKRNDESTFY]-x-[GREHFY]-C-[LIVM]	47	47	2	2		3 (3)
(7) Interleukin_18	(1) [EQ]-[SY]-S-[SL]-x(2)-[GS]-x-[FY]-L-[AST]-[CF]	41	41	3	0		
(8) Receptivity factor	(1) L-[LIVMPAG]-x(2)-[YF]-[LIVM]-x(2)-[QLIVM]-[GA]-x-P-[LIVMFY]-x-[DENHKRLIVM]-[PAG]-[DEAGST]-[FY]	204	204	0	0		
(9) GMP-beta	(1) [FY]-[LIVM](2)-x-[STAG]-[FYWH]-x(5)-[DE]-x(5)-P-[LIVM]-x(2)-[LIVM]-[FYWN]-x(2)-P	29	29	1	0	Q9VJL6; Q29NM1	9 (12)
Hormones							
(10) ACTH_domain and opioids neuropeptides	(1) K-R-[YF]-G-G-F-[LIVMT]-[STGKRIV]-[AGKRSTLIVMPY]; (2) K-R-[YF]-G-G-F-[LIVMT]>; (3) K-[KN]-[YF]-G-G-F-M-[KR]; (4) <[YF]-G-G-F-[LIVMT]-[STGKRIV]-[AGKRSTLIVMPY]; (5){CFYWHM}-Y-x-[MIVSTFY]-[FY]-H-F-R-W; (6) <Y-x-[MIVSTFY]-[FY]-H-F-R-W	397	1045	0	4		
(11) FMRFamide and related neuropeptides	(1) {LCFY}-[LCFYQWST]-[LCFYQWH]-[LCDEFYKRQW]-[LVMI]-[MLIV]-R-F-G-K-R; (2) {LCFY}-[LCFYQWSTLIVM]-[LCFYQWHKR]-[LCDEFYKRQWLIVM]-[LM]-[MIV]-R-F-G-R-[ASPD]-[LCFYHKR]-[LCQST]; (3) <x(0,8)-[LVMI]-[MLIV]-R-F>; (4) {CLIVM}-[CAGLIVMW]-[QCFYLW]-[FY]-[MLIV]-R-F-G-K-R; (5) {CHIV}-x-[CQN]-{HIV}-[CLIVMY]-{CAGLIVMW}-[QCFYLWIV]-[FY]-[MLIV]-R-F-G-R-[DNESTAG]; (6) <x(0,9)-[FY]-[MLIV]-R-F>; (7) {AGED}-[LIVMFY]-Q-G-R-F-G-R-[DEN]; (8) P-[AGST]-[LIVM]-R-[MLIV]-R-F>; (9) N-Q-[VI]-R-F-G-K-R; (10) [STG]-[LVMI]-F-R-F-G-K-R; (11) [RD]-[QPH]-F-[FY]-R-F-G-[KR]-[FWYL]; (12) [RD]-[QPH]-P-[FY]-R-F>; (13) R-P-[VI]-G-R-F-G-[KR]-[RS]; (14) S-A-[LM]-A-R-F-G-[KR]-[RS]; (15) [PQ]-[HL]-[LMFY]-R-G-R-F-G-R; (16) [STNFYH]-[LQ]-P-Q-R-F-G-[KR]-[LC]; (17) F-M-[NH]-F-G-K-R; (18) [AGNQ]-[GLE]-P-[LI]-R-F-G-[KR]-{QLIVMAG}; (19) P-[RK]-P-L-R-F-G>; (20) [FL]-G-T-M-R-F-G-[KR]-[RS]; (21) Q-[WL]-[LMIV]-[AGKRST]-G-R-F-G-[KR]; (22) [GA]-[GA]-[FY]-[ST]-[FY]-R-F-G-[RK]; (23) [GA]-[GA]-F-[ST]-[FY]-R-F>	214	605	0	2	Q7YWT6; Q622X3; Q51P51; Q616K2; Q613X6; Q21656; P34405; Q60ZQ9; Q618S3; Q620F8; Q620P9; Q7PUD4; Q618T6; Q705J7; Q3SXL4; Q3KNG4; Q60YH4; Q622X1; Q28Z02; Q297C5; Q28Z02	

(12) Neuropeptide-like protein*	(1) G-M-Y-G-G-[FYW]-G-R; (2) A-Q-[FW]-G-Y-G-[GY]-x(2)-[KRFYG]; (3) G-[FYW]-G-G-Y-G-G-Y-G-R-G; (4) P-L-Q-F-G-K-R; (5) [STRIV]-M-S-F-G-K-R; (6) [AGIV]-M-[AG]-F-G-K-R; (7) [DE]-K-R-G-G-A-R-A-[FYLIVM]; (8) R-x-G-[FML]-R-P-G-K-R; (9) [RFYM]-[AGTR]-F-A-F-A-K-R	33	84	0	7	Q60NA1; Q619H9; Q624T4; Q61BN3; Q62715; Q60MJ8; Q625G9; Q622L1; Q622L2
(13) Wamide neuropeptides*	(1) [QRKED]-[P]-[KRPQN]-[IVP]-G-[LM]-W-G-R-[RDESA]; (2) [ANPRKQ]-x-[AGLQP]-[RHKLIVP]-G-[LM]-W-G-K-R; (3) K-[KR]-x(1,5)-W-x(6)-W-G-[KR]-R	10	86	0	1	Q7Q4X3; Q8T3G1; Q60TK2; Q2LZG9
(14) Thyroliberin	(1)[KR]-[HKR]-Q-H-P-G-[KR]-R	12	78	1	0	
(15) Neurotensin/neuromedin N	(1) [KR]-[IVTRK]-P-Y-I-L-K-R; (2) [KR]-[IVTRK]-P-Y-I-L>	14	24	0	0	
(16)Allatostatin* (most of proteins are not characterized by existing motifs)	(1) [KR]-R-[NCKRFY]-x(0,11)-[FY]-[DENAGST]-[FY]-G-[LIVM]-G-[KR]-R; (2) <x(0,11)-[FY]-[DENAGST]-[FY]-G-[LIVM]>; (3) [KR]-R-x(0,3)-[FY]-[DENAGST]-[FY]-G-[LIVM]>	52	222	1	2	Q7QAG2; Q29BZ8
(17) Egg-laying hormone	(1) K-R-R-[LIVM]-R-F-[HNY]-[KR]-R; (2) P-R-[LIVM]-R-F-[HNY]-[PSTDEN]-x-[KRG]-[KR]-[KR]; (3) P-R-[LIVM]-R-F-[HNY]-[PSTDEN]-x(1,2)>	21	32	2	0	
(18) Periviscerokinin	(1) <x(0,1)-[AG]-x(0,3)-[GS]-[LIVM]-[LIFY]-x-[FYAMV]-[AGPM]-R-x>	59	59	0	0	
(19) Somatostatin	(1) C-[KRM]-[NSIV]-[FY]-[FY]-W-[KRDE]-[STG]-x-[ST]-x-C	71	71	2	0	
(20) Orcokinin*	(1) [KR]-R-N-F-[DE]-[DE]-[IV]-[DE]-[KR]; (2) <N-F-[DE]-[DE]-[IV]-[DE]-[KR]	3	22	0	0	Q7Q025; Q7QNH4; Q9W1F8; Q292P8
(21) Allatotropin*	(1) N-x(4)-[STIV]-A-R-G-[FY]-G-[KR]-R; (2) N-x(4)-[STIV]-A-R-G-[FY]>	15	18	0	1	Q7QKW9; Q7PZX1
(22) Ghrelin and Motilin-related peptide	(1) G-[STL]-[ST]-F-[LIVM]-[ST]-P-x(0,1)-[AGSTDE]-[FYQHM]-[QRK]; (2) [FY]-[VILM]-P-x-[FY]-[TS]-x(2)-[DE]-[LIVM]-[QRK]-[RK]-x-[QRK]-[ED]-[KR]	68	68	12	0	
(23) ADM	(1) [AG]-C-[P]-x-[AGFY]-[STMLIV]-C-[AGQIVT]-[VMLIFYHKR]-[QH]-x-[LIVM]	23	23	1	0	Q4RDH7; Q6IFS9
(24) Hepcidin* (most of proteins are not characterized by existing motifs)	(1) C-[CGW]-x-C-C-[C](4,5)-[CG]-G-x-C-C	44	44	0	1	Q4RUL1; Q4RUL2
(25) Achatin*	(1) K-R-G-F-[AGF]-[DG]-K-R; (2) <G-F-[AGF]-[DG]>	5	20	0	0	
(26) Cocaine- and amphetamine-regulated transcript protein	(1) C-x-C-x(5)-C-x(3)-[LIVM]-L-K-[C>]	11	11	2	0	Q4RMR3; Q568S2; Q68EU1; Q4SGG2; Q4T695; Q4TB13;
(27)Bradykinin	(1) P-[PAT]-G-[FW]-[ST]-P-[FL]-R	58	84	6	1	Q5XJ76
(28) GBP/PSP1/paralytic	(1) N-[FY]-x(2)-[GA]-C-x(2)-[GA]-[FY]-x-[RK]-[TS]-x-[DE]-[GA]-[RK]-C-[KR]-x-[TS]	18	18	0	0	
(29) Stanniocalcin	(1) C-L-x(2,6)-[GA]-C-x(2,5)-F-x-C-x(4)-[ST]-[CS]	45	45	1	0	
(30) Resistin	(1) C-x-C-x(3)-C-x(2)-W-x(7)-C-x-C-x-C-x(4)-W-x(4)-C-C	22	22	2	0	
(31) Pro-MCH	(1) [RK]-R-x(2,6)-[LMIV]-x-C-[MLIV](2)-[GA]-[RK]-[VLIM]-[FY]-x(2)-C-W; (2) R-[ED]-x(2)-[DE](3)-N-[ST]-[AG]-x-[FY]-[FK]-[IV]-[GD]-[RK]-R	29	39	4	0	
(32) Pigment dispersing hormone	(1) K-R-N-[ST]-[DEGA]-[LIVM](2)-N-[STAG]-[LIVM](2); (2) <N-[ST]-[DEGA]-[LIVM](2)-N-[STAG]-[LIVM](2)	21	21	0	1	Q298P6
(33) Orexin	(1) [HQ]-A-A-G-[IV]-L-T-[LIVM]-G-[KR]-R; (2) [HQ]-A-A-G-[IV]-L-T-[LIVM]>	11	18	0	0	
(34) Leucokinin*	(1) [PQAGSTKRH]-x-F-[HYN]-[AGSP]-W-[GA]-G-K-R; (2) <x-[PQAGSTKRH]-x-F-[HYN]-[AGSP]-W-[GA]>	11	11	0	0	Q60MR3; Q8MNU5
(35)Myomodulin*	(1) [LIVM]-[HQPST]-M-L-R-L-G-K-R	3	29	0	0	

Table 2 *(Continued)*

Name	Pattern	Tr. Po.	Hi.	Fa. Ne. Fr.	Fa. Ne. Pr.	New	Non-metazoa
(36)Nitrophorin	(1) C-[ST]-x(9,10)-[KRH]-x(2)-[FYW](2)-x(3,4)-[FYW](2)-x-[TS]-x-[FY]-x(4,5)-[PTS]	11	11	0	1		
(37)Prokineticin	(1) Q-C-x(4)-[CFY]-C-x(2)-[ST]-x(3)-[KR]-x-[LIVM]-[RK]-x-C-x-P-x-[GA]-x(2)-[GA]-x(2)-C-[HYF]-P	35	35	1	0		
(38) Leptin	(1)L-x-[VIT]-[FY]-[QRH]-[QKA]-[IV]-[LIVMH]-x-[SNG]-[LM]-[PHQS]	68	68	9	4		
Antimicrobial							
(39)Bombinin	(1) K-R-[LIVM](2)-G-P-[LIVM](2)-x(2)-[VILM]-[STG]-x(2)-[LIVM]-x(2)-[LIVM](2); (2) <[LIVM](2)-G-P-[LIVM](2)-x(2)-[VILM]-[STG]-x(2)-[LIVM]-x(2)-[LIVM](2); (3) [SG]-I-G-x(0,3)-[LIV]-x(2,7)-K-[STAGIV]-[AGFYIV]-[LIVF]-[KR]-[GAC]-[AGFYL]-[AGLVIM]-[KRN]	59	110	0	0		
(40) Brevinin, Dermaseptin, Aurein, Caeridin, Caerin, Dahlein, Temporin Ponericin and Uperin	(1) <x(7)-[C](2)-x(0,68)-C-[KSTAGLVE]-[LIVA]-[STAKYD]-[KRYGN]-[KRDESTQLG]-C>; (2) C-[KSTAGLVE]-[LIVA]-[STAKYD]-[KRYGN]-[KRDESTQLG]-C-R-x>; (3) <[DGA]-[LIVF]-[LIVMFW]-[DNESAGQKPLM]-[STLIVMKFAGDN]-[LVIMAGTF]-[KRAGSTVIL]-[KRHDENGASTQ]-[LIVMAGKFYSTW]-[IVLMAGFKRH]-[AGKRHSTDENQLIV]-[W]-x(0,2)>; (4) <[DGA]-[LIVF]-[LIVMFW]-[DNESAGQKPLM]-[STLIVMKFAGDN]-[LVIMAGTF]-[KRAGSTVIL]-[KRHDENGASTQ]-[LIVMAGKFYSTW]-[IVLMAGFKRH]-[AGKRHSTDENQLIV]-[W]-[CP](2)-x(0,35)>; (5) <x(0,45)-[QAGR]-[FYLQKRST]-K-R-[DGA]-[LIVFW]-[LIVMFW]-[DNESAGQKPLFM]-[STLIVMKFAGDN]-[LVIMAGTFY]-[KRAGSTVIL]-[KRHDENGASTQ]-[LIVMAGKFYSTW]-[IVLMAGFYKRH]-[AGKRHSTDENQLIV]-[W]-x(0,37)>; (6) <x(0,1)-[FIVLM]-[LIVMFYST]-[PGAQ]-x-[LIVMFY]-[AGSTIVLM]-[KRSTNDEMLIV]-[LIVMAGFY](0,1)-[LIVMAG](0,1)-x(0,2)-[GKRDEST]-[LIVM](2)>; (7) K-R-[FIVLM]-[LIVMFYST]-[PGAQ]-x-[LIVMFY]-[AGSTIVLM]-[KRSTNDEMLIV]-[LIVMAGFY](0,1)-[LIVMAG](0,1)-x(0,2)-[GKRDEST]-[LIVM](2)-G-K>	278	310	5	20		1 (1)
(41) Dermorphin	(1) K-R-Y-A-F-x-[YVLI]-[PVILM]-x-[RG]; (2) <Y-A-F-x-[YVLI]-[PVILM]-x>	6	22	0	0		
(42)Termicin*	(1) C-x(4)-C-W-x(2)-C-x(12)-C-x(4)-C-x-C	21	21	0	0		
(43)Liver-expressed antimicrobial	(1) [KR]-P-x(4)-C-x(5)-C-x(3)-[LIVM]-C-[KR]-x(2)-[RKHQ]-[CQ]	15	15	0	0	Q4SXZ9; Q5M9I7	
(44) Penaeidin	(1) [CR]-x(1,3)-C-[C](2)-[LIVM]-[C](7)-[CYF]-[CST]-[C](3)-[GA]-x-C-C	40	40	0	0		
(45) Ceratotoxin*	(1) [ST]-[LIVM]-[GA]-[ST]-[AG]-x-[KR]-[KR]-[AG]-[LIVM]-P-[LIVM]-[AG]-[KR](2)	10	10	3	0		
(46)Attacin	(1) [GTS]-[AGVMLI]-[AGFYST](0,1)-[FYLIV]-[AGDEL]-[GMQWKRHNDE]-[PKR]-[NKG]-[ADENHIV](0,1)-[NDEKR](0,1)-[GSR]-[HFL]-[GAS]-[GAL]-[STAED]-[LIVM]-[TSMQ]-[KRHDNEGA]-[TSEAG]-[HKRQGT] (2) Y-x-Q-[KRH]-L-[FG]-G-P-Y-G-N-S-x-P	50	50	0	1	Q290V6; Q291C0; Q295K8; Q29QF8; Q29QG5	

(47)Beta-defensin	(1) <x(0,79)-[WP]-x-C-[C]-[CP]-[CW]-[CA]-[C](0,4)-C-[CP]-[C]-[CW]-[C](0,2)-C-[C](3)-[CP](2)-[C](2)-[CP]-[C](1,5)-C-[C](0,3)-[C](4)-C-C-[CDENFWYP]-x(0,128)>	326	326	3	10	Q32P86; Q2XXN6; Q2XXN7; Q2XXN8; Q2XXN9	
(48) 4 kDa defensin*	(1) G-[CGA]-P-x(2)-[HQP]-x(2)-[CRK]-[DE]-x-[HP]-[CRWK]-[KR]-G-[MLIVEDN]	27	27	0	0		
Toxin							
(49)Conotoxin scaffold III/IV, mu-conotoxin and M conotoxin	(1) <x(0,62)-[C]-x(2)-[C](10)-C-C-[C](2,6)-C-[C](2,5)-C-[C](1,5)-C-[C](0,3)-C-[C](0,3)>; (2) <[C](0,9)-C-C-[C](2,6)-C-[C](2,5)-C-[C](1,5)-C-[C](0,1)-C-[C](0,3)>	62	62	0	0		
(50)Conotoxin scaffold IX and tau conotoxin	(1) <x(0,49)-[C](12)-[CDEFY]-[C](2)-C-C-[C](4,7)-C-[C](0,2)-C-[C](0,9)>; (2) <[C](0,14)-C-C-[C](4,7)-C-[C](0,2)-C-[C](0,9)>	80	80	0	1		
(51)Conotoxin scaffold VI/VII, four-loop conotoxin, Spider potassium channel inhibitory toxin, O superfamily	(1) <x-[PA]-x(0,17)-[C](0,21)-[C](2)-[CQ]-[C](11)-[CI]-[CP]-[C]-[CH]-C-[C](3,6)-C-[QC]-[C](3,9)-C-C-[C](2,8)-C-[CQ]-[C](2,9)-C-[C](0,9)>; (2) <[C](0,16)-[CQ]-C-[CI]-[C](2,5)-C-[QPC]-[C]-[CY](2)-[C](0,6)-C-C-[C](2,8)-C-[CQ]-[C](2,9)-C-[C](0,9)>; (3) <C-[CI]-[C](2,5)-C-[QPC]-[C]-[CY](2)-[C](0,6)-C-C-[C](2,8)-C-[CQ]-[C](2,9)-C-[C](0,9)>	408	408	2	23		11 (15)
(52) Scorpion toxin	(1) [CKDEN]-[C](3)-[CI]-[CDEN]-[C](2)-C-[C](3)-C-[C](6,10)-G-[C](1,2)-[CF]-x-[C](3,11)-C-[WYF]-C; (2) [CKDEN]-[C](3)-[CI]-[CDEN]-[C](4,9)-C-[C](3)-C-[C](6,10)-G-[C](1,2)-[CF]-x-[C](3,11)-C-[WYF]-C	223	223	9	5	Q2TSD9	
(53) Scorpion short toxin 2	(1) C-x-P-C-x(10)-C-x(2)-C-C-x(5,7)-C-x(2,3)-Q-C-[LIVM]-C	14	14	0	0		
(54) Anenome neurotoxin	(1) C-x-C-[C](4)-P-x(6,8)-G-x(5,13)-C-x(6,9)-C-x(6,9)-C-C	25	25	0	0		
(55) Melittin	(1) [LIVM]-[GA]-x(2)-[LIVM]-[KR]-[LIVM](2)-x(3)-[LIVM]-P-x-[LIVM](2)-x-W-[LIVM]	11	11	0	0		

Tr. Po.: the number of true positive peptide or precursor proteins; Hi.: the number of matches to the pattern, and there may be more than a hit to the pattern within a protein; Fa. Ne. Fr.: the number of false negative protein fragments; Fa. Ne. Pr.: the number of false negative peptide precursor proteins; New: the novel putative peptide precursors belonging to the family predicted by the corresponding pattern; non-metazoa: the number of proteins in non-metazoa which match the peptide patterns. The number of all known non-metazoan proteins, which have similar molecular function to the metazoan proteins in the corresponding peptide family, is also listed in parentheses.

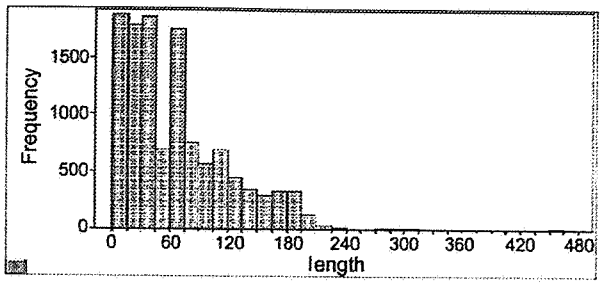

Fig. 2 – Histogram of peptide length distribution.

part. For some families of peptides, we have found additional novel patterns and these 34 novel patterns are marked as 'new' in Table 1.

In addition, 121 peptide patterns were newly identified. These patterns are listed in Table 2. These patterns allow the identification of 55 peptide families that are dismissed by PROSITE signatures. In total, these novel patterns currently cover 3866 bioactive peptides, cleaved from 3572 peptide precursors.

Of all the peptide patterns (Table 2), 28 patterns represent 12 peptide families that are not characterized by any other motif database, such as Pfam and Smart. These families are indicated with '*' and include the Orcokinin family and Wamide neuropeptide family. The patterns reminiscent for these families are short, occur repeatedly within the same protein sequence. The sequences outside the conserved region are not well preserved, and thus a probability model based on protein sequence alignments cannot efficiently characterize such peptide families.

4.3. *Case study*

Patterns respectively representing the family of opioid and POMC-derived peptides and the FMRFamide related neuropeptides are here shown as test cases in order to provide insight in how the peptide patterns perform.

4.3.1. *First case*

The family of *Opioid neuropeptides*. Opioid peptides are neuropeptides that are involved in pain control mechanisms

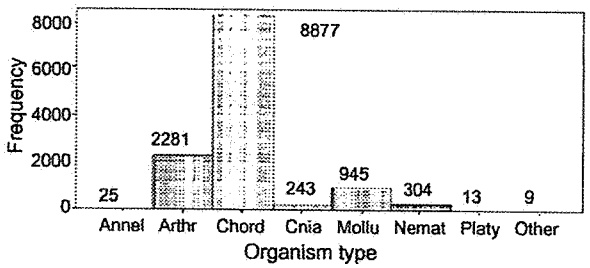

Fig. 3 – Histogram of peptide distribution among phyla. Annel: Annelida; Arthr: Arthropoda; Chord: Chordata; Cnia: Cnidaria; Mollu: Mollusca; Nemat: Nematoda; Platy: Platyhelminthes; Other: consisting of species *Echinodermata* and *Echiura* containing seven and two peptides, respectively.

in vertebrates. The PROSITE pattern PS01252 'endogenous opioids neuropeptides precursors' (C-x(3)-C-x(2)-C-x(2)-[KRH]-x(6,7)-[LIF]-[DNS]-x(3)-C-x-[LIVM]-[EQ]-C-[EQ]-x(8)-W-x(2)-C) covers 39 UniProtKB/TrEMBL entries that match the pattern. These 39 proteins include proenkephalin (PENK), nociceptin (PNOC) and prodynorphin (PDYN) peptide precursors. However, 92 remaining proteins that are a PENK, PNOC or PDYN peptide precursor cannot be identified by the PROSITE pattern, including nine full peptide precursors such as Q7T3L0 from *Zebrafish*, and 83 peptides or partially sequenced fragments such as Q5G6A0 from *Nycteris grandis* and human Q9BYY3.

The PROSITE pattern for opioid peptides is the only PROSITE pattern that is not identified by our search procedure. This can be explained as follows: the regions of the protein sequences matching the PROSITE pattern are outside the sequence areas covering the bioactive peptides. Our search program detects conserved regions within the mature bioactive peptide sequences. Nevertheless, our procedure identifies six novel peptide patterns instead. These new patterns are not only reminiscent of all PENK, PNOC and PDYN peptides and precursor proteins, but also of proopiomelanocortin (POMC) peptide precursors, for which no PROSITE motif exists. POMC peptide precursors are a sister family of the PENK, PNOC and PDYN peptide family. POMC is known to share similar peptide sequences with PENK, PNOC and PDYN precursors, but also contains other non-opioid peptide sequences such as ACTH and α-MSH, which are involved in the stress response and stimulate corticosteroid release [1].

In total, 113 peptide precursors were found to contain two of the six novel peptide patterns, and 284 other ones match one of the patterns. These patterns characterize conserved domains located at different regions of a precursor, and each of them can exclusively represent an opioid peptide or peptide precursor protein family.

Apart from PROSITE, there are other motif databases that have identified motifs for peptide families. The conserved domain database (CDD) [13] is a collection of multiple sequence alignments for ancient domains and full-length proteins, including domains imported from SMART, Pfam and COGs [18]. Conserved motifs are created based on the multiple sequence alignments in the form of Position Specific Scoring Matrix (PSSM). Two motifs in the CDD database represent the family of opioid and POMC derived peptides, i.e. 'ACTH_domain' (PF00976) and 'Vertebrate endogenous opioids neuropeptide' (PF01160). They are 41 and 243 columns (amino acids) in length respectively, and are imported from Pfam which contains two additional motifs characterizing this peptide family including 'opioids neuropeptide' (PF08035, 31 amino acids) and 'Pro-opiomelanocortin, N-terminal region' (PF08384, 45 amino acids). When querying the CDD database at the NCBI website (http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml) using default parameters, we found that 27 proteins in the family of opioid and POMC derived peptides, mostly peptide precursor sequence fragments, cannot be identified by any of the two motifs with a significant score (for example, *e*-value <0.01) (for example PENK Q9W687 from *Acipenser transmontanus* and horse beta-melanotropin P01202). When we further searched for these peptides in the Pfam database itself (http://www.sanger.ac.uk/Software/Pfam/

search.shtml) by means of the fragment search mode (fs), none of these 27 proteins that matches one of the 4 Pfam motifs in this family with bit scores higher than gathering threshold could be detected.

4.3.2. *Second case*

The family of *FMRFamide and related neuropeptides (FARPs)*. It is widely known that FARPs occur throughout the whole animal kingdom and therefore this family is an ideally suited test case to check whether the identified pattern is capable of retrieving FARPs throughout all metazoan species where these peptides have been sequenced. In total, we could identify 214 FARPs using the presently identified FARP peptide patterns, and these FARPs show 605 hits to the patterns due to the presence of multiple copies of the conserved regions within several FARP precursor proteins. The identified FARPs distribute among a wide range of phyla, including Nematoda (85), Arthropoda (50), Mollusca (24), Annelida (9), Platyhelminthes (1), Cnidaria (10) and Chordata (35). The Clustal-W multi-alignment of all these FARP sequences together or within each of the seven phyla using default parameters (http://www.ebi.ac.uk/clustalw/) shows that the FARP protein precursors display sequence similarities within the mature peptide regions, particularly in the area containing the conserved peptide patterns, and that the remaining parts of the precursor sequences display rather low similarities. The FARP peptide precursors also differ from each other by the number of peptide repeat units. The difference in the copy number of peptides is thought to have arisen by unequal crossover events [11]. We also observed that most of the mature FARP peptides in Nematoda, Mollusca, Arthropoda, Annelida and Platyhelminthes share common C-terminal sequences such as F-[LVMIFY]-R-Famide, Y-[MLIV]-R-Famide and [LVMI]-[LVMIF]-R-Famide, but have different N-terminal extensions.

The PROSITE database does not contain a signature representing the FARP family of peptides and the CCD database contains an 11-amino-acid motif PF01581 that characterizes FARPs from six phyla including Nematoda, Mollusca, Arthropoda, Annelida, Platyhelminthes and Cnidaria. FARPs from Chordata are not retrieved by this CCD motif, for example FMRFamide-related peptide precursors Q9HCQ7 from *Human* and Q9WVA8 from *Mouse*. In addition, conversely to the present peptide motif database, 49 FARP peptides or precursor proteins in these above-mentioned six phyla (for example, the FARP peptide Q9TWD2 from *Lymnaea stagnalis* and a FARP peptide precursor Q95QP2 from *C. elegans*) cannot be identified by the CCD motif with a significant score (*e*-value <0.01). The CDD motif, which is based on the alignments of intact precursor proteins, is not sufficient to identify all members of the FARP peptide family.

5. Discussion

Protein domains are highly conserved throughout evolution and there are several databases available that catalogue protein families and domains. Such motif and domain databases are very useful in assigning a putative function to an unannotated protein. Peptide precursor proteins are a special class of proteins because they undergo extensive post-translational processing before producing final mature peptides making the annotation of peptides and peptide precursor proteins challenging. This is illustrated by the fact that many metazoan peptides and peptide precursors are not represented by the motifs currently present in the widely used motif database such as the PROSITE database.

Because of the broad functional range of motifs available to date and because of the wide range of peptide (precursor protein) families, a comprehensive database of conserved patterns typical for endogenously occurring mature peptides will be of great value. We have designed a searching procedure to find conserved patterns within each of the known peptide families, and as a result, we have constructed a motif database comprising 211 patterns that are representatives most currently known peptide families. In total, 121 novel peptide patterns are identified as shown in Table 2 covering 3866 bioactive peptides or 3572 peptide precursor proteins in total. These novel peptide patterns represent 55 peptide families which can not be identified by the currently available PROSITE signatures. The peptide pattern for a particular peptide family in our peptide pattern database often represents a conserved region different from the pattern reminiscent of the same family in PROSITE. For instance, the PROSITE pattern for the somatotropin peptide family matches 483 proteins, while a newly identified pattern 'C-[LIVMFG]-x-[KHRSNDEQVI]-[DEN]-{CNDEPQ}-{AGLMVI}-[KRMT]-{DENKRHPQ}-x-[STNALIVMF]-[FYLIVMKS]-[LIMVT]-x-{NDEKRH}-[LIVMATE]-[KRNEQTA]-C' covers 522 proteins belonging to this peptide family.

Many peptides have been isolated and sequenced as mature peptides and their precursor proteins are often unknown as yet. Therefore, these small peptides are difficult to be identified by CCD or other motif databases. Most of the conserved motifs in CDD database are long ones attempting to feature the full-length protein sequences, and thus they can efficiently identify the proteins which have overall protein sequence similarities. Motifs in databases such as Pfam contain two Hidden Marcov Models (HMMs) for each family based on a multiple protein sequence alignment, one built to find complete domains (ls mode) and the other to match fragments of domains (fs mode) [5]. These motifs are sensitive at identifying complete domains and thus they can efficiently identify the proteins which have sequence similarities that cover the full length protein sequence or at least contain a complete domain. However, these motifs do not work very well when they encounter short peptides which lack information on amino acids at the sites outside the peptide sequences, or when the conserved regions are limited, especially in distantly related proteins where the overall-length sequence similarity may be not well preserved. For example, in the family of Opioid and POMC-derived peptides, the six novel peptide patterns represent not only the peptide precursor molecules but also the mature fully processed peptides.

Conservative peptide sequence patterns correspond to functionally and structurally important parts of the peptide, i.e. the binding site to specific receptors, the disulphide bonds for stability and tertiary structure. The identification of a peptide motif in proteins will be very useful in designing experiments to test the function of specific proteins, predicting the structure of proteins, and identifying new members of protein families. For example, scanning the peptide patterns

against Uniprot protein database revealed 95 proteins (listed in Tables 1 and 2) which are as yet unannotated as putative peptide or precursor proteins.

Our pattern search procedure can also be applied to detect conserved patterns in other peptide families which are not annotated by peptide keywords, such as accessory gland-specific peptides and immune-induced peptides.

When determining short functional patterns for peptide sequences, we have to evaluate whether all peptides or peptide precursor proteins in the 110 characterized peptide families are represented by the motifs and whether false positives are identified by the motifs. Short motifs often have some degree of degeneracy and the presence of a motif in a protein may reflect a conserved functional role, a yet to be discovered structural functional role or a non-functional role. When using the short currently identified peptide patterns, while the false positives are kept to zero, we observe that 440 (3.8%) of the mature peptides or sequence fragments and 282 (2.5%) of the peptide precursor proteins in these 110 characterized families cannot be identified by these peptide patterns. Many of them could be determined by combining the peptide pattern search procedure with the structural hallmarks of bioactive peptides and their precursors [12], such as the length of a peptide precursor which is usually not longer than 500 amino acids, the presence of a signal peptide which directs a precursor protein into the secretary pathway of the cell, and the presence of typical cleavage sites flanking the mature peptide [12]. To be even more successful in identifying all false negatives while eliminating all false positives because of the short length and degeneracy of most short motifs, it may be possible to make use of 3D structural patterns when they become available for peptide precursor proteins. Patterns that integrate 3D structural information of the pattern sequences will be more sensitive in identifying the peptide precursors [7,19].

While the majority of known peptide families have been characterized by the established peptide patterns, a few peptide families could be missed due to the lack of annotation information on the members of these corresponding families. In addition, 297 remaining peptides from 251 precursor proteins are not processed by the pattern search procedure. They are from small peptide families, such as eclosion hormones, ecdysis-triggering hormones and apelin, which have only a few homologies so far. A pattern based on the small number of peptides usually cannot gain enough confidence in representing the subfamily, and also cannot sufficiently reflect the sequence divergence formed in the evolutionary course of the family member.

Although our initial training peptide datasets exclusively consisted of metazoan bioactive peptides, we find it interesting to note that for 14 out of 110 metazoan peptide families, identified in this study, the corresponding peptide patterns also retrieves peptides from other kingdoms such as protoctists, fungi and plants. These non-metazoan peptides include defensins, antibacterial peptides and growth factors. Numerous growth factors including FGF, NGF, PDGF, and chemokines were also retrieved from viral genomes. Although one might immediately classify these proteins as false positives, further evaluation using CLUSTALW alignments (http://www.ebi.ac.uk/clustalw/index.html) and careful anlysis of GO annotations reveal that molecular functions similar to their metazoan counterparts have been electronically annotated to these viral proteins. On the other hand, the mining result is not surprising as these viral signalling molecules are probably the result of evolution of recognition systems by the virus. Viruses identify their host cells by a lock-and key fit between viral proteins and specific receptor molecules on the surface of their host cells. It is therefore no surprise that viral proteins resemble the host's own signalling molecules to ensure specific recognition of host cells, and the viral genes that encode mimics of growth factors, cytokines and chemokines may have been pirated from the host genome during the long co-evolution of virus and host. These findings are in agreement with the recent identification of viral proteins with homology to cellular proteins such as chemokines and chemokine receptor-like chemokinin G-protein coupled receptors in Cytomegalovirus [20].

The ability to retrieve functionally annotated peptides from resources, not included in the initial training set further validates the presently identified peptide patterns. Tables 1 and 2 list the number of proteins in non-metazoa that match the peptide patterns. The number of all known non-metazoan proteins sharing a similar molecular function to the metazoan proteins in the corresponding family is also listed in the following bracket.

In sum, the present peptide pattern database is widely applicable for the identification of critical functional residues in proteins and for annotation of hypothetical proteins using the novel motifs assembled in our peptide motif database.

Acknowledgement

This research was sponsored by the FWO grant G0146.03.

REFERENCES

[1] Arends RJ, Vermeer H, Martens GJ, Leunissen JA, Wendelaar Bonga SE, Flik G. Cloning and expression of two proopiomelanocortin mRNAs in the common carp (*Cyprinus carpio* L.). Mol Cell Endocrinol 1998;143(1–2):23–31.

[2] Baggerman G, Liu F, Wets G, Schoofs L. Bioinformatic analysis of Peptide precursor proteins. Ann N Y Acad Sci 2005;1040:59–65.

[3] Bateman A, Birney E, Cerruti L, Durbin R, Etwiller L, Eddy SR, et al. The Pfam protein families database. Nucl Acids Res 2002;30(1):276–80.

[4] Bateman A, Coin L, Durbin R, Finn RD, Hollich V, Griffiths-Jones S, et al. The Pfam protein families database. Nucl Acids Res 2004;32(Database issue):D138–41.

[5] Durbin R, Eddy S, Krogh A, Mitchison G. Biological sequence analysis: probabilistic models of proteins and nucleic acids Cambridge University Press; 1998.

[6] Filipsson K, Kvist-Reimer M, Ahren B. The neuropeptide pituitary adenylate cyclase-activating polypeptide and islet function. Diabetes 2001;50(9):1959–69.

[7] Gribskov M, Homyak M, Edenfield J, Eisenberg D. Profile scanning for three-dimensional structural patterns in protein sequences. Comput Appl Biosci 1988;4(1):61–6.

[8] Henry J, Favrel P, Boucaud-Camou E. Isolation and identification of a novel Ala-Pro-Gly-Trp-amide-related

peptide inhibiting the motility of the mature oviduct in the cuttlefish, *Sepia officinalis*. Peptides 1997;18(10):1469–74.
[9] Hulo N, Sigrist CJ, Le SV, Langendijk-Genevaux PS, Bordoli L, Gattiker A, et al. Recent improvements to the PROSITE database. Nucl Acids Res 2004;32(Database issue):D134–7.
[10] Jonassen I, Collins JF, Higgins DG. Finding flexible patterns in unaligned protein sequences. Protein Sci 1995;4(8): 1587–95.
[11] Lee HS, Simon JA, Lis JT. Structure and expression of ubiquitin genes of *Drosophila melanogaster*. Mol Cell Biol 1988;8(11):4727–35.
[12] Liu F, Baggerman G, D'Hertog W, Verleyen P, Schoofs L, Wets G. In silico identification of new secretory peptide genes in *Drosophila melanogaster*. Mol Cell Proteomics 2006;5(3):510–22.
[13] Marchler-Bauer A, Anderson JB, Cherukuri PF, Weese-Scott C, Geer LY, Gwadz M, et al. CDD: a conserved domain database for protein classification. Nucl Acids Res 2005;33(Database issue):D192–6.
[14] Masashi Y, Watanobe H, Terano A. Central regulation of hepatic function by neuropeptides. J Gastroenterol 2001;36:361–7.
[15] Rouille Y, Duguay SJ, Lund K, Furuta M, Gong Q, Lipkind G, et al. Proteolytic processing mechanisms in the biosynthesis of neuroendocrine peptides: the subtilisin-like proprotein convertases. Front Neuroendocrinol 1995; 322–61.
[16] Schlesinger DH, Pickart L, Thaler MM. Growth-modulating serum tripeptide is glycyl-histidyl-lysine. Experientia 1977;33(3):324–5.
[17] Schultz J, Milpetz F, Bork P, Ponting CP. SMART, a simple modular architecture research tool: identification of signaling domains. Proc Natl Acad Sci USA 1998;95(11):5857–64.
[18] Tatusov RL, Galperin MY, Natale DA, Koonin EV. The COG database: a tool for genome-scale analysis of protein functions and evolution. Nucl Acids Res 2000;28(1): 33–6.
[19] Taylor WR, Jonassen I. A structural pattern-based method for protein fold recognition. Proteins 2004;56(2): 222–34.
[20] van Cleef KW, Smit MJ, Bruggeman CA, Vink C. Cytomegalovirus-encoded homologs of G protein-coupled receptors and chemokines. J Clin Virol 2006;35(3):343–8.
[21] Vandenborne K, Roelens SA, Darras VM, Kuhn ER, Van der GS. Cloning and hypothalamic distribution of the chicken thyrotropin-releasing hormone precursor cDNA. J Endocrinol 2005;186(2):387–96.