6

A Pattern Search Method for Discovering Conserved Motifs in Bioactive Peptide Families

Feng Liu[1], Liliane Schoofs[2], Geert Baggerman[2],
Geert Wets[1] and Marleen Lindemans[2]
[1]*Data Analysis & Modeling Group, Transportation Research Institute, Hasselt University*
[2]*Functional Genomics and Proteomics, Department of Biology, K.U. Leuven*
Belgium

1. Introduction

Bioactive peptides play critical roles in regulating most biological processes in animals, and they have considerable biological, medical and industrial importance. Peptides belonging to the same family are often characterized by a typical short sequence motif (pattern) that is highly functionally preserved among the family members. In this chapter, we design a pattern search method to facilitate the detection of such conserved motifs. First, all known bioactive peptides annotated in Uniprot are collected and classified, and the program Pratt is used to search these unaligned peptide sequences in each family for conserved patterns. The obtained patterns are then refined by taking into account the information on amino acids at important functional sites collected from literature, and are further tested by scanning them against all the Uniprot proteins. The diagnostic power of the patterns is demonstrated by the fact that, while the false positive is kept to zero to ensure that the signatures are exclusive to peptides and their precursors, nearly 94% of all known peptide family members accommodate one or several of the identified patterns.
In total, we brought to light 155 novel peptide patterns in addition to the 56 established ones in the PROSITE database. All the patterns represent 110 peptide families; among which 55 are not characterized by PROSITE and 12 are also dismissed by other existing motif databases, such as Pfam. Using the newly uncovered peptide patterns as a search tool, we predicted 95 hypothetical proteins as putative peptides or peptide precursors.

2. Problem statement and background

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 1.7 million protein sequences, which are currently available for all the completely sequenced metazoan genomes, nearly 15% could not be assigned to any putative function. Although several tools/algorithms are available to contribute towards the putative functional assignments of the proteins, yet large numbers of proteins remain un-elucidated. In most cases this is due to the low degrees of sequence similarities with known proteins; alternatively, the existing similarities can be confined to only very small part(s) of the entire protein. The latter is especially true for precursor proteins coding for bioactive peptides. Consequently, there is still a need for

bioinformatic tools to predict the function of the enormously large number of the unknown protein sequences.

Bioactive peptides occur in the whole animal kingdom, from the least evolved phyla to the highest vertebrates (Filipsson et al., 2001; Masashi et al., 2001). They play key roles as signaling 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 (Boonen et al., 2007; Boonen et al., 2010). They are synthesized in the cell in the form of large preproproteins (precursors), which are a special class of proteins as they undergo extensive post-translational processing prior to producing final mature bioactive peptides (Schoofs & Baggerman, 2003). Peptides and their precursors that are structurally and functionally related have been classified into peptide families; each family of proteins is assumed to be derived from a common ancestor (Husson et al., 2009). During the evolutionary process, the protein sequences may have much diverged, but the essential amino acids involved in the biologically important activities are still present. These conserved amino acids along with their particular sequential order form the functional foundation and represent the motif (pattern) of a peptide family.

However, over the course of natural adaptation, different peptide families have diverged at different rates. While for some peptide families, the similarity extends over a much longer region even over the entire peptide precursor sequences; for many others, a short highly conserved motif is responsible for the function of the precursor proteins throughout the family members, and the sequence fragments outside the conserved regions often display no significant similarities (Baggerman et al., 2005). The latter conserved sequence characteristics can be further exposed by many short but biologically important functional peptides released from known large precursors as annotated in Uniprot, such as the 3-amino-acid thyroliberin peptide 'QHPamide' (Vandenborne et al., 2005) and 4-amino-acid neuropeptides 'FMRFamide' (Baggerman et al., 2002). For some mature peptides, the precursor proteins (genes) are unknown, such as the 2-amino-acid neuropeptide 'GWamide' (P83570) from *Sepia officinalis* (Henry et al., 1997) and the human growth-modulating peptide 'GHK' (P01157) (Schlesinger et al., 1977). The existence of numerous short bioactive peptides within the precursor proteins implies that only a very small conserved peptide motif may be a biologically important functional portion of the precursors.

Due to the fact that only short sequence regions are conserved, peptides or their precursors are sometimes not identified by existing sequence alignment algorithms e.g. BLAST or by motif search methods. While BLAST programs (Altschul et al., 1997) are very suitable to scan databases for homologous proteins, they are far less efficient at finding similarities to short conserved regions which can be only a few amino acids in length, when the whole genome sequence is scanned. For large precursors which are usually a few hundred amino acids in length and for which the biologically conserved regions are limited, the important domains are often masked by long randomly unrelated sequence regions. This is because for any two random large protein sequences, BLAST usually can find a relative long local alignment, at least longer than the short conserved peptide motif, and BLAST tends to assign a higher score to a longer alignment (Durbin et al., 1998). In addition, if a pair of homologues involves a short independent peptide molecule, which may be either an unknown peptide sequence as query or a known mature peptide as target from a protein database, it is difficult for BLAST to detect the pair of homologues, because the involvement of a short sequence makes the pairwise sequence alignment less likely to obtain a significant BLAST score (e.g., e-value < 0.01).

Like BLAST, motif search methods are important tools to search for a protein in a database, nevertheless, they are also limited to detect all members from a characterized peptide family. Most of the motifs in the existing databases, e.g. PROSITE (Hulo et al., 2004) and Pfam (Finn et al, 2010), cover the entire precursor sequences or sequence domains which are much longer than the conserved bioactive peptide regions. Therefore, the database motifs show their weakness when they are used to detect short mature peptides for which the precursors are unknown and the information on the sequences outside the peptide regions is thus missing. In addition, the construction of these motifs requires a good multiple protein sequence alignment in order to produce an accurate signature. This works well when the sequences are easy to align. However, for some peptide families for which the conserved regions are very short and the bulk of peptide precursor sequences is not very well preserved, the multiple alignment is very difficult to obtain or evaluate. The overall precursor protein sequence identity, especially in distantly related homologues, may be too low for an accurate alignment. In some cases, the short conserved regions are repeated within a precursor, making it even more challenging to build a unique alignment that truly reflects the evolutionary relationship.

In this chapter, 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 peptides and peptide precursor proteins in Metazoa through a search in Uniprot and classified them into peptide families. Next, we extracted peptide sequences in each family and used the program Pratt to search the sequences for representative patterns. Such patterns consist of highly conserved positions that can be separated by fixed or variable spacing. The patterns are then refined by incorporating the information that is available in literature on the important amino acids contained within the biologically active site(s) of the peptides. The specificity of the generated patterns are further verified by scanning them against Uniprot in order to ascertain that proteins picked up by the patterns are either annotated as peptides or peptide precursor proteins or have an unknown function.

3. Data collection

3.1 Peptide precursor collection and classification

A protein was collected into a peptide-precursor database if it is annotated in the Uniprot protein database (release 6.6) consisting of Swiss-Prot (release 48.6) and TrEMBL (release 31.6) with one of the following keywords: hormone, antimicrobial, toxin. The hormone includes 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 this database). The antimicrobial consists of antibiotic, antiviral defense, defensin and fungicide; while the toxin includes naturally produced and secreted poisonous proteins that damage or kill other cells. However, when the protein is also characterized by non-peptide keywords, such as receptor, signal-anchor, transmembrane, binding protein, DNA binding, nuclear protein, transport, collagen, enzyme or words ending in 'ase' (excluding 'disease'), it is excluded, in order to avoid the selection of proteins which are not peptides or peptide precursors.

Stand-alone PSI-BLAST (ftp://ftp.ncbi.nih.gov/blast/executables/) is then used to align all the assembled sequences with all the Uniprot proteins except the ones which are already in

the peptide-precursor database. Based on the conserved sequence characteristics of peptide families, the score matrix PAM30 is used and the word size is set to 2, allowing for the search for short but strong similarities. The proteins, which show significant similarities (e-value <0.01) with the known peptides or precursors, are retained. The obtained list is then checked manually in terms of the proteins' cellular location, molecular function and biological process as stated by GO (gene ontology) terms or in literature. As a result, 1345 more proteins which have as yet not been annotated in Uniprot are added to the peptide-precursor database.

Proteins collected in this database are automatically classified into peptide families if their family classification information is available in Uniprot that is based on a significant match to an existing motif or based on sequence similarities. Otherwise, proteins that display sequence similarities with a significant BLAST score, are clustered into the same family. A protein can also be assigned to a particular family based on its molecular function described in literature.

3.2 In silicon extraction of peptides

From each precursor protein in a peptide family, the bioactive peptide sequences are extracted in silicon from the beginning and ending positions of the subsequences that are annotated as 'peptide' or 'chain' in 'feature' line in the corresponding protein file in Uniprot. The conserved basic cleavage sites flanking the peptides, which contribute to the endoproteolytic cleavage process of the peptides from their precursors, such as the monobasic site (G)R or (G)K, the dibasic sites (G)KR, (G)RR, (G)KK or (G)RK, or a combination of consecutive K or R, are also withdrawn along with the subsequences (Liu & Wets, 2005; Rouille et al., 1995).

Entries in the family that only constitute the peptide sequence, i.e. in those cases where the precursor is unknown, are also retained. Proteins less than 200aa (amino acids) in length, which contain an N-terminal signal peptide and for which no mature peptides have as yet been identified, presumably contain a single peptide and are therefore also deposited after in silicon removal of the N-terminal signal peptide. According to the statistics on all annotated bioactive peptide sequences in Uniprot, 97% are no longer than the 200aa threshold value. The presence of a signal peptide is assumed when it is indicated in Uniprot; in other cases, it is forecasted by the signal peptide prediction program signalP (http://www.cbs.dtu.dk/services/SignalP/).

In total, 110 datasets of peptide families are formed with each including at least 10 peptide sequences. All the extracted peptide sequences in each of the families were scanned independently for patterns conserved in the corresponding family.

4. Method

Different software available on the internet provides users the tools to search for patterns conserved in a set of unaligned protein sequences. Pratt (http://www.ebi.ac.uk/pratt/#) (Jonassen et al., 1995) is a flexible pattern search tool in the number of parameters that can be controlled by users. It allows searching 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 peptide family datasets, and the searching parameters are set based on maximum pattern length and pattern flexibilities found in the existing peptide patterns in PROSITE.

For each Pratt run which starts with the minimum percentage of sequences to match the pattern (the parameter C%) equal to 90%, the most significant pattern, which is the one with the highest fitness in the Pratt output list, is retained. The obtained pattern is then refined by integrating the information on the important functional sites in the matched peptide sequences depicted in literature. The amino acids occurring at these sites are added to the pattern if they are absent at the corresponding sites in the pattern.

The pattern is further verified by scanning it against all the Uniprot proteins 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-peptide protein, it is retained as a conserved peptide pattern. (2) Otherwise, if the pattern is matched by both peptide and non-peptide proteins (further referred to as true and false positive hits, respectively), it is subsequently processed as follows. (2a)If the pattern does not include any wildcard region where any amino acid is accepted, the positions where the pattern is located in all matching protein sequences are checked. If the pattern exclusively occurs at the N- or C-terminus of the true positive hits, or if the peptide proteins are all 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 and the N- or C-terminus of the peptide or precursor protein. If the pattern with such a restriction cannot distinguish the true positives from the false ones, the pattern is eliminated. (2b)Or, if the pattern has wildcard regions, the sequence fragments corresponding to the pattern in all the matching sequences are extracted and aligned. If the two groups of amino acids in a wildcard region X in this alignment have different physicochemical properties between the true and the false positive hits, the region X is replaced by the group of amino acids distinctively occurring in the true positive proteins. In the other case, when the two groups of amino acids share identical physicochemical properties, the pattern is discarded. The amino acid symbol sets: DE, KRH, NQ, ST, ILV, FWY, AG, C, M and P, which are classified based on the physiochemical nature of the side groups (Smith & Smith), are used.

If a conserved pattern cannot be obtained, the parameter C% is reduced by 10%, and Pratt is re-run against the same dataset. As the percentage of sequences to match the pattern decreases, a pattern which is usually longer and contains more sites than the previously one is shown up and processed by similar refinement and verification. The procedure is repeated until a pattern, which represents the majority of a group of related peptide sequences and rules out 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 this dataset. The sequence regions which match the pattern are removed from the original peptides. Each of the two remaining parts of the peptide sequences at their N- and C-terminus is left to form an independent sequence if it is not less than 4aa 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 that match the pattern. This methodology is based on the fact that a peptide precursor protein may contain several conserved regions, and that our extracted peptide sequences include 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 until the parameter C% is less than 50%.

This means that the remaining dataset contains no more patterns representing the majority of the sequences.
Fig. 1 represents the scheme of the described pattern searching procedure which is aimed to examine 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 literature, but also the general information which is extracted by the computational tool Pratt from all related peptides in a family.

5. Results

5.1 'PeptideMotif' database

We have built a peptide-precursor database consisting of 11,688 peptides and precursor proteins originated from 1420 metazoan organisms; of which 11,437 proteins (98%) are categorized into 110 distinctive peptide families. Based on bioactive peptide sequences drawn from the peptide families, we uncovered in total 211 conserved patterns which are assembled into the peptide motif database 'PeptideMotif'.
All the patterns range between 4 and 52 amino acids (column) in length with 78 (37%) no longer than 10aa. While each of the patterns covers most of the peptides or precursors belonging to the corresponding family, the false positives are kept to zero because it is guaranteed by the criterion that a known protein matching the pattern is indeed a peptide or precursor protein from this family.

5.2 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 56 entries (patterns) describing 55 peptide families in Metazoa (the omega-atracotoxin family has two patterns) belonging to categories of cytokines and growth factors, hormones and active peptides, and toxins. All the 55 families are also covered by patterns in the 'PeptideMotif' database, and these peptide patterns (Table 1) share the similar length to their PROSITE counterparts. However, in terms of conserved sequence characteristics revealed in both database motifs, more amino acids are imposed at the conserved sites or wildcard regions in the 'PeptideMotif' patterns. This is due to the fact that the identified peptide patterns are not only trained by running them against the Swiss-Prot protein database which is also used as the test dataset by PROSITE, but also against the TrEMBL database, in which many proteins are also annotated by keywords or literature. In addition, for 25 of the 56 families, we have found 34 additional novel patterns and they are marked as 'new' in Table 1.
The remaining 121 'PeptideMotif' patterns presented in Table 2 allow the identification of 55 peptide families that are untouched by PROSITE signatures; they cover 3866 bioactive peptide sequences cleaved from 3572 precursors. Among the patterns, 28 representing 12 families are also not characterized by any other motif database, such as Pfam (Bateman et al., 2004) and CDD (Marchler-Bauer et al., 2005). The sequence reminiscence for these families is short and often occurs repeatedly within a same precursor protein. 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.

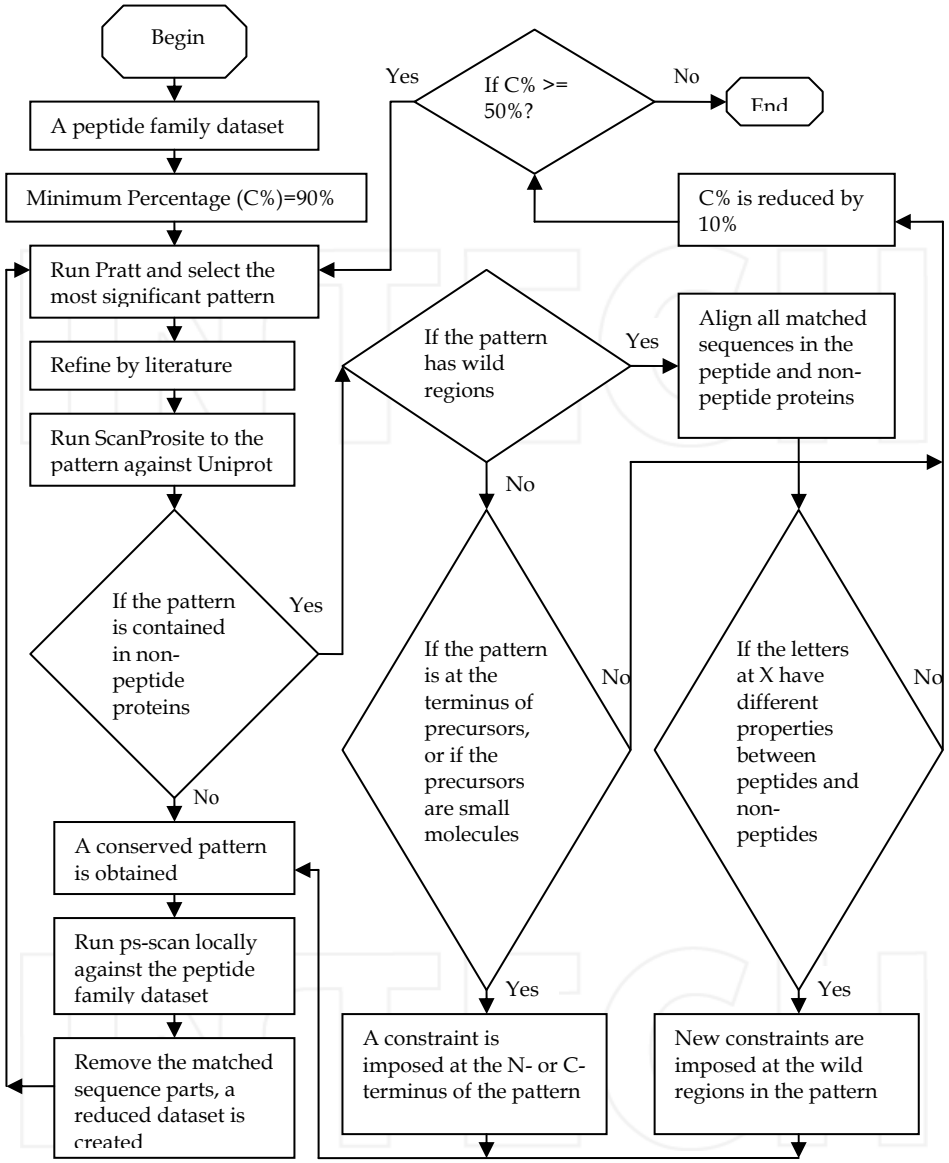


Fig. 1. Procedure for searching patterns in peptide sequences.
Note: The parameters are set as follows: the maximum pattern length (PL) is 52, the maximum length of a wildcard (PX) is 15, the maximum number of flexible wildcards (FN) is 3, the maximum flexibility of a flexible wild card (FL) is 8, the upper limit on the product of flexibilities for a pattern (FP) is 48, the minimum percentage of sequences to match the pattern (C%) is 90, 80, 70, 60 and 50%, respectively, and all other parameters are at default.

Cytokines and growth factors
(1) Granulins; (1) C-x-D-x(2)-H-C-C-{LIVM}-x(4)-C; {42, 241, 2}; {Q616A1, Q7JKP2, Q9U362}
(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); {300,530,44}
(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}
(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, 12}
(5)Platelet-derived growth factor (PDGF); (1) P-[PSRAKQGL]-C-[LIVMFYAGST]-x(3)-[RQ]-C-[AGSTMLIVN]-G-S(0,1)-[CN]-C; {158, 158, 23}
(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, 18}; { Q6DUZ6, Q6GLX8, Q4T8B9}
(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, 234,27}; { Q3ZBN3, Q32L58}
(8)TGF-beta; (1) [WFYSTKRHL]-[LIVM]-[LIVMKRHF]-{CPNL}-P-{FY}-{PCW}-[FYILVA]-{C}-{QCWKRH}-{PA}-{PAGC}-C-{C}-[GE]-{C}-C; {766, 766, 59}
(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, 29}
(10)Granulocyte-macrophage colony-stimulating factor; (1) C-P-[LP]-T-{ST}-E-x-{QLIVMT}-C; {25, 25, 1}; {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, 24}
(12) Interleukin_2; (1) [ST]-E-[LF]-x(2)-L-x-C-L-x-[EDN]-E-L; {74, 74, 14}
(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]-x(11,17)-C (new); {73, 119, 4}
(14) Interleukin_6; (1) C-x(9)-C-[FYLIVM]-x(5)-G-L-x(2)-[FY]-x(3)-L; {69, 69, 8}
(15) Interleukin_7_9; (1) N-[DAT]-[LAPS]-[SCT]-F-L-K-{AGDE}-L-L; {20, 20, 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}
(17) LIF / OSM; (1) [PSTA]-x(4)-F-[NQ]-x-K-x(3)-[CG]-x-[LF]-L-x(2)-Y-[HK] ; {24, 24, 4}
(18) Osteopontin; (1) P-x(1,5)-[KQ]-x-[TA]-x(2)-[GA]-S-S-E-E-K; {27, 27, 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} {Q5TTQ9}
(20) Bombesin-like peptides (1) [HLIVMQ]-W-A-[STIVRK]-G-[SH]-[LF]-M; {42, 42, 1}
(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}

(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, 9}; {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} {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, 5}; {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, 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, 4}
(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, 8}
(29) Glycoprotein hormones alpha chain (1) C-x-G-C-C-[FY]-S-x-A-[FY]-P-T-P; {109, 109, 4}
(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, 13}
(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}
(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, 52} {Q32L79, Q621L6, Q61VN2, Q61GN7, Q4T1R8}
(33) Natriuretic peptides (1) C-F-G-x(3)-[DEA]-[RH]-I-x(3)-[ST]-x(2)-G-C; {155, 155,10}
(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}
(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}
(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, 7}
(37) Parathyroid hormone (1) [KR]-R-x-[VI]-[STAGFYN]-[EH]-x-Q-x(2)-H-[DEN]-x-[GR]; {54, 54, 3}
(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, 4} {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, 45}
(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, 124,6}

(41)Urotensin II (1) C-F-W-K-Y-C (identical); {30, 30,1}
(42) Endothelin (1) C-{C}-C-{C}(4)-D-{C}(2)-C-{C}(2)-[FY]-C; {50, 104, 2}
(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}
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, 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, 5}
(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, 7}; {Q6XD83}
(47) Cathelicidins (1) Y-{LIVM}-[EDQN]-[AVI]-[LMVI]-{HKRG}-[RKHQ]-A-[LIVMA]-[DQGEN]-x-[LIVMFY]-N-[DEQ]; {58, 58, 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, 20}
(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}
(50) Scorpion short toxin 1 (1) C-{C}(4,5)-C-{PC}-{CQ}-{C}-C-x(3)-{C}-{CPWA}-x(1,4)-[GASEDN]-[KRAVISNDE]-C-[VIMQTDK]-[NG]-x(1,2)-{P}-C-[HKRDENVI]-C; {77, 77, 6}
(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, 1}
(52) I-superfamily conotoxin (1) C-{C}(6)-C-{C}(5)-C-C-{C}(1,3)-C-C-{C}(2,4)-C-{C}(3,10)-C (identical); {37, 37, 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, 2}
(54) Omega-atracotoxin (ACTX) (1)C-[IT]-P-S-G-Q-P-C (identical); (2)C-C-[GE]-[ML]-T-P-x-C (identical); {13, 13, 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}

Table 1. The conserved peptide patterns similar to PROSITE signatures.

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, 44}
(2) Interleukin_3 (1) [CVLIM]-[LIVM]-P-x-[AGPST]-x(2)-[STAGDENRKH]-x(12,14)-[DE]-F-[RKQ]-{NDEAGQST}-K-L; {20, 20, 0}
(3) Interleukin_5 (1) [HDE]-x(2)-C-x(3)-[IVLM]-F-x-G-[LIVMST]-x(2)-L-x-[NST]; {23, 23, 1}
(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}
(5) Interleukin_15(1)C-{C}(4)-[LM]-{C}-C-[FY]-[LIVFYQ]-x-[DE]-[LIVM]-x(2)-[LIVM]-x(2)-[ED]; {44, 44, 1}

(6) Interleukin_17 (1) [RLM]-{QKR}-[PS]-{P}-x-[LIVMFY]-{RKH}-{CP}-[AS]-x-Cx-[CHKRNDESTFY]-x-[GRKHFY]-C-[LIVM]; {47, 47, 4}
(7) Interleukin_18 (1) [EQ]-[SY]-S-[SL]-x(2)-[GS]-x-[FY]-L-[AST]-[CF]; {41, 41, 3}
(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}
(9) GMF-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}; {Q9VJL6, Q29NM1}
Hormones
(10) ACTH_domain and opioid 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, 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-GR-[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]-F-[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]-PQ-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>; {214,605,2}; {Q7YWT6, Q622X3,Q61P51, 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-GK-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-PG-K-R; (9) [RFYM]-[AGTR]-F-A-F-A-K-R; {33, 84, 7}; {Q60NA1, Q619H9, Q624T4, Q61BN3, Q627I5, 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, 1} {Q7Q4X3, Q8T3G1, Q60TK2, Q2LZG9}
(14) Thyroliberin (1)[KR]-[HKR]-Q-H-P-G-[KR]-R; {12, 78, 1}
(15) Neurotensin/neuromedin N (1)[KR]-[IVTRK]-P-Y-I-L-K-R; (2) [KR]-[IVTRK]-P-Y-I-L>; {14, 24, 0}
(16)Allatostatin* (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, 3}; {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}
(18)Periviscerokinin (1)<x(0,1)-[AG]-x(0,3)-[GS]-[LIVM]-[LIFY]-x-[FYAMV]-[AGPM]-R-x>;{59, 59, 0}
(19) Somatostatin (1) C-[KRM]-[NSIV]-[FY]-[FY]-W-[KRDE]-[STG]-x-[ST]-x-C; {71, 71, 2}
(20) Orcokinin* (1) [KR]-R-N-F-[DE]-[DE]-[IV]-[DE]-[KR]; (2) <N-F-[DE]-[DE]-[IV]-[DE]-[KR]; {3, 22, 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, 1}; {Q7QKW9, Q7PZX1}

(22) Ghrelin and Motilinrelated 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}
(23) ADM (1) [AG]-C-{P}-x-[AGFY]-[STMLIV]-C-[AGQIVT]-[VMLIFYHKR]-[QH]-x-[LIVM]; {23, 23, 1}; {Q4RDH7, Q6IFS9}
(24) Hepcidin* (1) C-[CGW]-x-C-C-{C}(4,5)-[CG]-G-x-C-C; {44, 44, 1}; {Q4RUL1, Q4RUL2}
(25) Achatin* (1) K-R-G-F-[AGF]-[DG]-K-R; (2) <G-F-[AGF]-[DG]>; {5, 20, 0}
(26) Cocaine- and amphetamineregulated transcript protein (1) C-x-C-x(5)-C-x(3)-[LIVM]-L-K-[C>]; {11, 11, 2}; {Q4RMR3, Q568S2, Q68EU1, Q4SGG2, Q4T695, Q4TBI3}
(27)Bradykinin (1) P-[PAT]-G-[FW]-[ST]-P-[FL]-R; {58, 84, 7}; {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}
(29) Stanniocalcin (1) C-L-x(2,6)-[GA]-C-x(2,5)-F-x-C-x(4)-[ST]-[CS]; {45, 45, 1}
(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}
(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]-[PK]-[IV]-[GD]-[RK]-R; {29, 39, 4}
(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, 1}; {Q298P6}
(33) Orexin (1) [HQ]-A-A-G-[IV]-L-T-[LIVM]-G-[KR]-R; (2) [HQ]-A-AG-[IV]-L-T-[LIVM]>; {11, 18, 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}; {Q60MR3, Q8MNU5}
(35)Myomodulin* (1) [LIVM]-[HQPST]-M-L-R-L-G-K-R; {3, 29, 0}
(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, 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}
(38) Leptin (1)L-x-[VIT]-[FY]-[QRH]-[QKA]-[IV]-[LIVMH]-x-[SNG]-[LM]-[PHQS]; {68, 68, 13}
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]-IG-x(0,3)-[LIV]-x(2,7)-K-[STAGIV]-[AGFYIV]-[LIVF]-[KR]-[GAC]-[AGFYL]-[AGLVIM]-[KRN]; {59, 110, 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, 25}
(41) Dermorphin (1) K-R-Y-A-F-x-[YVLI]-[PVILM]-x-[RG];(2) <Y-A-F-x-[YVLI]-[PVILM]-x>; {6, 22, 0}
(42)Termicin* (1) C-x(4)-C-W-x(2)-C-x(12)-C-x(4)-C-x-C; {21, 21, 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}; {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}
(45) Ceratotoxin* (1) [ST]-[LIVM]-[GA]-[ST]-[AG]-x-[KR]-[KR]-[AG]-[LIVM]-P-[LIVM]-[AG]-[KR](2); {10, 10, 3}
(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-[PG]-G-P-Y-G-N-S-x-P; {50, 50, 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, 13}; {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}
Toxin
(49)Conotoxin scaffold III/IV, muconotoxin 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}
(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, 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, 25}
(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, 14}; {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}
(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}
(55) Melittin (1) [LIVM]-[GA]-x(2)-[LIVM]-[KR]-[LIVM]; (2)-x(3)-[LIVM]-P-x-[LIVM](2)-x-W-[LIVM]; {11, 11, 0}

Table 2. The novel conserved peptide patterns.
Note: each family is described in the following items: (1) the name of the family; (2) all identified patterns; patterns marked with 'identical' are completely identical to their PROSITE counterpart and the ones marked as 'new' are novel to PROSITE in Table 1; (3) the number of true positive peptide or precursor proteins, the number of matches to the pattern, and the number of false negative hits, all these numbers are in a bracket; (4) if there are novel putative peptides or precursors predicted by the patterns of the family, they are listed in a second bracket.

6. Case study

Patterns respectively representing the family of opioid and POMC-derived peptides as well as the FMRFamide and related neuropeptides (FARPs) are here shown as test cases in order to provide insights into the conserved sequence characteristics in many know peptide families and how the peptide patterns deduced based on these characteristics perform.

6.1 Opioid and POMC-derived peptides

The family includes subfamilies of opioid peptides and pro-opiomelanocortin (POMC) proteins, and proteins in this family vary in length ranging from large precursors with a few hundred amino acids, e.g. Q805B5 in *Chimaera phantasma* (325aa), to short peptides or partial sequence fragments, e.g. Q7M2Z6 in Sheep (13aa).

6.1.1 The subfamily of opioid peptides

Opioid peptides are neuropeptides that are involved in pain control mechanisms in vertebrates, and they consist of proenkephalin (PENK), nociceptin (PNOC) and prodynorphin (PDYN) (Comb et al., 1982). The 41-column PROSITE pattern PS01252 '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' matches 39 Uniprot proteins. However, 92 remaining sequences from the subfamily are disregarded; including nine full peptide precursors e.g. zebrafish Q7T3L0 and 83 peptides or sequence fragments e.g. human Q9BYY3.

The subfamily is also described by a 71-column Pfam motif PF01160. When querying this motif against all proteins in the subfamily by means of 'both global (ls) and fragment (fs)' search modes (http://www.sanger.ac.uk/Software/Pfam/search.shtml), 78 precursors are singled out. But, the other 53 opioid proteins, e.g. cat Q28409, zebrafish Q8AX66 and Q9W687 from *Acipenser transmontanus*, cannot be recognized by the Pfam motif with a score higher than a gathering threshold.

A further investigation into the proteins missed by the Pfam motif is conducted by comparing them with all proteins in the non-redundant protein sequence database nr using BLAST (http://www.ncbi.nlm.nih.gov/BLAST). The alignments with Q28409 (Fig. 2) reveal that, while the similarities between the two Mammal precursors Q28409 and P01210 are conserved along the entire sequences, the resemblances between Q28409 and Q8AX66/Q4RIZ7 from the remote phylum of Actinopterygii are confined to a limited region identified as '[KR]-[KR]-Y-G-G-F-[ML]-[KR]-[KR]'. The few highly conserved amino acids are also observed from the alignments between Q9W687 and Q5Y3C6 from Chondrichthyes and Q6SYA7 from Dipnoi (Fig. 3). However, this conserved region is too short to produce a significant score, and therefore BLAST comparison alone will fail to detect the limited similarity preserved among the distant homologues with a critical confidence level.

The existing PROSITE pattern and the Pfam motif both characterize only the conserved N-terminal region of the peptide precursors, they are thus not sufficient in identifying all short bioactive opioid peptides or sequence fragments which are cleaved from their large precursors and do not carry the N-terminal part of the proteins, but nevertheless bring the crucial conserved peptide sequence region with them and preserve the fundamental function of the peptide subfamily. Therefore, although the sequences, e.g. Q28409, Q8AX66 and Q9W687, cannot be identified by the existing motifs, they all share the pattern '[KR]-[KR]-Y-G-G-F-[ML]-[KR]-[KR]' from our 'PeptideMotif' database. The pattern, which is derived from the bioactive peptide sequences, could be more functionally conserved and more performable in identifying opioid peptides or entire precursor proteins.

6.1.2 The subfamily of POMC-derived peptides

The subfamily shares similar peptide sequences with opioid precursors, but also contains other non-opioid peptides such as ACTH and alpha-MSH, which are involved in the stress response and stimulate corticosteroid release (Arends et al., 1998).

```
Query=Q28409|PENK_FELCA Proenkephalin A-Felis silvestris catus(Mammalia) Length=187

> P01210|PENK_HUMAN  Proenkephalin A precursor - Homo sapiens (Mammalia)
Length=267   Score =  429 bits (1004),  Expect = 1e-118

Query  WETCKEFLKLSQLEIPQDGTSALRESS-PEESHALRKKYGGFMKRYGGFMKKMDELYPQE
       WETCKE L+LS+ E+PQDGTS LRE+S PEESH L K+YGGFMKRYGGFMKKMDELYP E
Sbjct  WETCKELLQLSKPELPQDGTSTLRENSKPEESHLLAKRYGGFMKRYGGFMKKMDELYPME

Query  PEEEAP-AEILAKRYGGFMKKDAEEEEDALASSSDLLKELLGPGETETAAAPRGR-----
       PEEEA  +EILAKRYGGFMKKDAEE+ D+LA+SSDLLKELL  G+       R R
Sbjct  PEEEANGSEILAKRYGGFMKKDAEED-DSLANSSDLLKELLETGDN------RERSHHQD

Query  ---DDEDVSKSHGGFMRALKGSPQLAQEAKMLQKRYGGFMRRVGRPEWWMDYQKRYGGFL
       ++E+VSK +GGFMR LK SPQL  EAK LQKRYGGFMRRVGRPEWWMDYQKRYGGFL
Sbjct  GSDNEEEVSKRYGGFMRGLKRSPQLEDEAKELQKRYGGFMRRVGRPEWWMDYQKRYGGFL

Query  KRFADSLPSDEEGESYS
       KRFA++LPSDEEGESYS
Sbjct  KRFAEALPSDEEGESYS

>Q8AX66|Q8AX66_BRARE Proenkephalin (Fragment) - Brachydanio rerio (Actinopterygii)
Length=216  Score =  140 bits (324),  Expect = 9e-32

Query  KKYGGFMKRYGGFMKKMDELYPQEPEEEAPAEILAKRYGGFMKKDAE----EEED-----
       KKYGGFMKR                      +E L KRYGGFMKK AE    E ED
Sbjct  KKYGGFMKR----------------------SESLIKRYGGFMKKAAEFYGLESEDVDQGR

Query  ALASSSDLLKELLGP-----GETETAAAPRGRDDED-VSKSHGGFMR-----ALKGSPQL
       A+ ++ D+  E+L       GE E AA  R +  E+  +K +GGFMR     AL
Sbjct  AILTNHDV--EMLANQVEADGEREEAALTRSKGGEEGTAKRYGGFMRRGGLYAL------

Query  AQEA--KMLQKRYGGFMRRVGRPEWWMDYQ--KRYGGFLKRFADSLPSDEEGE
         E+  + LQKRYGGFMRRVGRP+WW   Q  KRYGGFLKR       S E+ E
Sbjct  --ESGVRELQKRYGGFMRRVGRPDWW---QESKRYGGFLKR------SQEQDE

> Q4RIZ7|Q4RIZ7_TETNG Chromosome undetermined SCAF15040 - Tetraodon nigroviridis
(Actinopterygii)  Length=246 Score =  123 bits (283),  Expect = 2e-26

Query  KKYGGFMKRYGGFMKKMD------ELYPQEPEEEA--PAEIL------------------
       KKYGGFMKRYGGFM + D      E +P +P+EE     EIL
Sbjct  KKYGGFMKRYGGFMSRRDVPEGALE-HPSDPDEEENIRLEILKILNAAAVHGSEGGGKAG

Query  --AKRYGGFMKKDAEEEEDALASSSDLLKELLGPGETETAAAPRGRDDEDVSKSHGGFMR
          KRYGGFM++ AEE     A+  DLL+ +LG                          R
Sbjct  EEGKRYGGFMRR-AEEG----AAQGDLLEAVLG--------------------------R

Query  ALKGSPQLAQEAKMLQKRYGGFMRRVGRPEW--------------WM---DYQKRYGGFL
        LK             KRYGGFMRRVGRPEW              W    D QKRYGGF+
Sbjct  GLK-------------KRYGGFMRRVGRPEWLVDSSKRGGVLKRAWGSDNDLQKRYGGFM
```

Fig. 2. Sequence alignments between Q28409 and P01210/Q8AX66/Q4RIZ7 by BLAST. Notes: the conserved opioid peptide sequence similarities are in bold.

No signature represents the subfamily in PROSITE; three Pfam motifs explain the proteins including PF08384 (45 columns), PF00976 (41 columns) and PF08035 (31 columns). These motifs capture separate conserved regions located respectively at the N-ternimus of the precursors after the removal of the signal peptide, at the sequences coding for ACTH and for 'beta-endorphin' peptides. However, the remaining parts of the precursors encoding for peptides of gamma-MSH (12aa) and beta-MSH (17aa) are left untouched. As a result, 27 mature peptides or sequence fragments, e.g. Q9PRN3 from the *Sea lamprey*, horse P01202 and leech P41989, cannot be detected by any of these Pfam motifs.

```
Query= Q9W687|Q9W687_ACITR Proenkephalin (Fragment)-Acipenser
transmontanus (Actinopterygii)   Length=45

> Q5Y3C6|Q5Y3C6_HETPO Proenkephalin - Heterodontus portusjacksoni
(Chondrichthyes)  Length=264 Score = 39.2 bits (85),  Expect = 0.032

Query  14   RYDGFSKQ------PEHTDSKEITSEEV---EKRYGGFM  43
            RY GF K+      P   D  EI S+EV   EKRYGGFM
Sbjct  225  RYGGFMKRWNDILVPSDEDG-EIYSKEVPELEKRYGGFM  262

Score = 31.2 bits (66),  Expect = 8.7

Query  14   RYDGFSKQPEHTDSKE--ITSEEVE----------KRYGGFM  43
            RY GF K+    DS +  I+  EV+          KRYGGFM
Sbjct  105  RYGGFMKK---ADSGDMYIS--EVDDENKGREILSKRYGGFM  141

> Q6SYA7|Q6SYA7_PROAN Prodynorphin (Fragment) - Protopterus annectens
(Dipnoi)   Length=191 Score = 33.7 bits (72),  Expect = 1.5

Query  33   EEVEKRYGGFM  43
            EE++KRYGGFM
Sbjct  169  EELQKRYGGFM  179
```

Fig. 3. Sequence alignments between Q9W687 and Q5Y3C6/Q6SYA7 by BLAST.
Note: the conserved opioid peptide sequence similarities are in bold.

The BLAST alignment between Q9PRN3 and all proteins in the nr database unveils that, although Q9PRN3 cannot be identified by the Pfam motifs, it shares the highly conserved 'PeptideMotif' pattern 'Y-x-[MV]-x-H-F-R-W' with other POMC subfamily members, e.g., Q2L6A9 from Hyperoartia, P01193 and Q53WY7 from Mammalia, and Q32U15 from Amphibia (Fig. 4). This 8-column peptide pattern is a part of the 41-column Pfam motif PF00976. While the sequence region, which is described by this Pfam motif, may be an entire functional or structural domain, this peptide pattern contained within the longer domain is probably the most essentially functional part.
In total, our procedure identifies six novel peptide patterns in the combination of these two subfamilies. Among all the 397 proteins in this family, 113 were found to contain two of the peptide patterns, and the rest match one of them. These patterns characterize conserved domains located at different regions of a precursor sequence, and each of them can exclusively represent an opioid or POMC peptide or its precursor protein.

6.2 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 disclosed pattern is capable of retrieving FARPs from all metazoan species (Ubuka et al., 2009). In total, 23 conserved peptide patterns have been uncovered from the family, and they match 214 FARPs sequences with 605 hits due to the presence of multiple copies of the conserved patterns within some 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).
An 11-column Pfam motif PF01581 characterizes FARPs from all above-mentioned phyla except Chordata, e.g. human Q9HCQ7 and mouse Q9WVA8. In addition, conversely to the 'PeptideMotif' patterns, 49 FARP peptides or precursor proteins in these characterized phyla, e.g., Q9TWD2 from *Lymnaea stagnalis* and Q95QP2 from *Caenorhabditis elegans*, cannot be revealed by the Pfam motif with a significant score (e-value <0.01).

```
Query= Q9PRN3|Q9PRN3_PETMA Melanotropin MSH-B - Petromyzon marinus
(Hyperoartia)  Length=20

>Q2L6A9|Q2L6A9_MORMR Proopiomelanotropin (Fragment) - Mordacia mordax
(Hyperoartia)  Length=154 Score = 51.5 bits (114),  Expect = 5e-06

Query  2   QESADGYRMQHFRWGQPLP  20
           QE+ D YR+QHFRWG+PLP
Sbjct  11  QENPDAYRIQHFRWGEPLP  29

> P01193|COLI_MOUSE Corticotropin-lipotropin precursor(Pro-
opiomelanocortin) (POMC) - Mus musculus (Mammalia)  Length=235
Score = 32.5 bits (69),  Expect = 2.4

Query  8    YRMQHFRWGQP  18
            Y M+HFRWG+P
Sbjct  125  YSMEHFRWGKP  135

Score = 30.8 bits (65),  Expect = 7.7

Query  3    ESADG-YRMQHFRWGQP  18
            E  DG YR++HFRW  P
Sbjct  183  EKDDGPYRVEHFRWSNP  199

Score = 22.3 bits (45),  Expect =  2753

Query  8   YRMQHFRW  15
           Y M HFRW
Sbjct  77  YVMGHFRW  84

> Q53WY7|Q53WY7_HUMAN Proopiomelanocortin (Fragment) - Homo sapiens
(Mammalia)  Length=30 Score = 22.3 bits (45),  Expect =  2753

Query  8   YRMQHFRW  15
           Y M HFRW
Sbjct  3   YVMGHFRW  10

> Q32U15|Q32U15_9NEOB Proopiomelanocortin A (Fragment) - Trachycephalus
jordani (Amphibia)  Length=82 Score = 23.1 bits (47),  Expect =  1529

Query  8   YRMQHFRW  15
           Y M HFRW
Sbjct  23  YVMSHFRW  30
```

Fig. 4. Sequence alignments between Q9PRN3 and P01193/Q53WY7/Q32U15 by BLAST. Note: the conserved peptide sequence similarities are in bold.

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 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 within the sequences, which is thought to have arisen by unequal crossover events (Lee et al., 1998). In addition, we also observed that most of the mature FARP peptides share common C-terminal sequences but have much mutated N-terminal extensions. All these make it problematic to construct an accurate multiple alignment in order to derive a statistical

model which represents distantly related proteins from various phyla throughout the evolutionary history of the FARP peptide family.

7. Conclusion

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 unknown protein. Peptide precursor proteins are a distinctive class of molecules because they undertake various posttranslational modifications in order to ultimately synthesize stabilized and functional 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 PROSITE.

Because of the tremendously increasing number of protein sequences and because of the wide range of peptide families, a comprehensive database of conserved patterns typical for endogenously occurring mature peptides is of great value in identifying new peptides and precursor proteins to catch up with their sequencing rate. We therefore have designed a searching procedure to find conserved patterns within the known peptides, and as a result, we have constructed a 'PeptideMotif' database that is representative of most currently known peptide families.

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 other motif databases. 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) (Durbin et al., 1998). These motifs are sensitive at identifying complete domains and thus they can efficiently detect the proteins which have 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. In contrast, the patterns derived directly from the mature peptide sequences grasp the highly preserved region of the precursor proteins and thus are able to identify not only the peptide precursor molecules but also the fully processed peptides.

Conservative peptide sequence patterns correspond to functionally and structurally important parts of the peptides, i.e. the binding site to specific receptors, the disulphide bonds for stability and tertiary structure. The discovery of peptide motifs will be undoubtedly of great value for any peptide-related studies ranging from the identification of putative peptides and precursor proteins to the annotation of critical functional residues (Husson et al., 2010), to the complement of peptidomic research in detecting and verifying peptides in vitro (Baggerman et al., 2004; Boonen et al., 2008; Menschaert et al., 2010). For example, scanning the peptide patterns against Uniprot revealed 95 proteins (listed in Tables 1 and 2) which are not as yet annotated as putative peptides or precursor proteins.

When determining short functional patterns for peptide sequences, we have to evaluate how representative the peptide motifs are in the 110 characterized peptide families. 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 described families cannot be recognized by the 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 (Liu et al., 2006), 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 peptides. 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 sequences will be more sensitive in identifying peptides and peptide precursors (Gribskov et al., 1988; Taylor et al., 2004).

While the majority of known peptide families have been profiled by the established peptide patterns, the remaining ones accounting for in total 251 peptides and precursor proteins (2% of all the proteins in the peptide-precursor database) 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 family, and also cannot sufficiently reflect the sequence divergence accumulated in the evolutionary course of the family member. As more peptides and precursor proteins are sequenced, our patterns search procedure can be applied to the corresponding families and the 'PeptideMotif' database will be updated accordingly, keeping the peptide pattern database widely applicable for the identification of critical functional residues and for the annotation of hypothetical molecules in various peptide families.

8. References

Altschul, S.F.; Madden, T.L.; Schaffer, A.A.; Zhang, J.; Zhang, Z.; Miller, W. & Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res,* Vol. 25, No. 17, pp. 3389-3402, ISSN 0305-1048

Arends, R.J.; Vermeer, H.; Martens, G.J.; Leunissen, J.A.; Wendelaar Bonga, S.E. & Flik G. (1998). Cloning and expression of two proopiomelanocortin mRNAs in the common carp (Cyprinus carpio L.). *Mol Cell Endocrinol,* Vol.143, No. 1-2, (August 1998), pp. 23–31, ISSN 0303-7207

Baggerman, G.; Cerstiaens, A.; De Loof, A. & Schoofs, L. (2002). Peptidomics of the larval Drosophila melanogaster central nervous system. *J. Biol. Chem,* Vol. 277, pp. 40368-40374, ISSN 0021-9258

Baggerman, G.; Liu, F.; Wets, G. & Schoofs, L. (2005). Bioinformatic analysis of peptide precursor proteins. *Ann N Y Acad Sci,* Vol.1040, pp. 59–65, ISSN 0077-8923

Baggerman, G.; Verleyen, P.; Clynen, E.; Huybrechts, J.; De Loof, A. & Schoofs, L. (2004). Peptidomics. *J Chromatogr B Analyt Technol Biomed Life Sci,* Vol. 803, No. 1, pp. 3-16, ISSN 1570-0232

Bateman, A.; Coin, L.; Durbin, R.; Finn R.D.; Hollich, V.; Griffiths-Jones, S.; Khanna, A.; Marshall, M.; Moxon, S.; Sonnhammer, E.L.L.; Studholme, D.J.; Yeats, C. & Eddy, S.R. (2004). The Pfam protein families database. *Nucl Acids Res,* Vol. 32, No. suppl 1, pp. D138–41, ISSN 0305-1048

Boonen, K.; Baggerman, G.; D'Hertog, W.; Husson, S.J.; Overbergh, L.; Mathieu, C. & Schoofs, L. (2007). Neuropeptides of the islets of Langerhans: a peptidomics study. *Gen Comp Endocrinol,* Vol. 152, No. 2-3, pp. 231-241, ISSN 0016-6480

Boonen, K.; Landuyt, B.; Baggerman, G.; Husson, S.J.; Huybrechts, J. & Schoofs, L. (2008). Peptidomics: the integrated approach of MS, hyphenated techniques and bioinformatics for neuropeptide analysis. *J Sep Sci,* Vol. 31, No. 3, pp. 427-445, ISSN 1615-9306

Boonen, K.; Husson, S.J.; Landuyt, B.; Baggerman, G.; Hayakawa, E.; Luyten, W.H. & Schoofs, L. (2010). Identification and relative quantification of neuropeptides from the endocrine tissues. *Methods Mol Biol,* Vol. 615, pp. 191-206, ISSN 1940-6029

Comb, M.; Seeburg, P.H.; Adelman, J.; Eiden, L. & Herbert, E. (1982). Primary structure of the human Met- and Leu-enkephalin precursor and its mRNA. *Nature*, Vol. 295, pp. 663-666, ISSN 0028-0836

Durbin, R.; Eddy, S.; Krogh, A. & Mitchison, G. (1998). *Biological sequence analysis: probabilistic models of proteins and nucleic acids,* Cambridge University Press, ISBN 9780521629713, Cambridge, UK

Filipsson, K.; Kvist-Reimer, M. & Ahren, B. (2001). The neuropeptide pituitary adenylate cyclase-activating polypeptide and islet function. *Diabetes,* Vol.50, No.9, pp. 1959–1969, ISSN 0012-1797

Finn, R.D.; Mistry, J.; Tate, J.; Coggill, P.; Heger, A.; Pollington, J.E.; Gavin, O.L.; Gunasekaran, P.; Ceric, G.; Forslund, K.; Holm, L.; Sonnhammer, E. L. L.; Eddy S. R. & Bateman A. (2010). The Pfam protein protein families database. *Nucleic Acids Res,* Vol.38, No. suppl 1, pp. D211-D222, ISSN 0305-1048

Gribskov, M.; Homyak, M.; Edenfield, J. & Eisenberg, D. (1988). Profile scanning for three-dimensional structural patterns in protein sequences. *Comput Appl Biosci,* Vol. 4, No. 1, pp. 61–66, ISSN 0266-7061

Henry, J.; Favrel, P. & Boucaud-Camou, E. (1997). 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,* Vol.18, No. 10, pp. 1469–1474, ISSN 0196-9781

Hulo, N.; Sigrist, C.J.; Le, S.V.; Langendijk-Genevaux, P.S.; Bordoli, L.; Gattiker, A.; De Castro, E.; Bucher, P. & Bairoch, A. (2004). Recent improvements to the PROSITE database. *Nucl Acids Res,* Vol. 32, pp. D134–D137, ISSN 0305-1048

Husson, S.J.; Clynen, E.; Boonen, K.; Janssen, T.; Lindemans, M.; Baggerman, G. & Schoofs, L. (2010). Approaches to identify endogenous peptides in the soil nematode Caenorhabditis elegans. *Methods Mol Biol,* Vol. 615, pp. 29-47, ISSN 1940-6029

Husson, S.J.; Landuyt, B.; Nys, T.; Baggerman, G.; Boonen, K.; Clynen, E.; Lindemans, M.; Janssen, T. & Schoofs, L. (2009) Comparative peptidomics of Caenorhabditis elegans versus C. briggsae by LC-MALDI-TOF MS. *Peptides,* Vol. 30, No. 3, pp. 449-457, ISSN 0196-9781

Jonassen, I.; Collins, J.F. & Higgins, D.G. (1995). Finding flexible patterns in unaligned protein sequences. *Protein Sci,* Vol. 4, No. 8, pp. 1587–1595, ISSN 0961-8368

Lee, H.S.; Simon, J.A. & Lis, J.T. (1998). Structure and expression of ubiquitin genes of Drosophila melanogaster. *Mol Cell Biol,* Vol. 8, No. 11, pp. 4727–4735, ISSN 0898-7750

Liu, F.; Baggerman, G.; D'Hertog, W.; Verleyen, P.; Schoofs, L. & Wets, G. (2006). In silico identification of new secretory peptide genes in Drosophila melanogaster. *Mol Cell Proteomics,* Vol. 5, No. 3, pp. 510–522, ISSN 1535-9476

Liu, F. & Wets, G. (2005). A Neural Network Method for Prediction of Proteolytic Cleavage Sites in Neuropeptide Precursors. *Proceedings of the 2005 IEEE Engineering in Medicine and Biology 27th Annual Conference,* Vol. 3, pp. 2805-2808, ISSN 1557-170X, ShangHai, China, September 1-4, 2005

Marchler-Bauer, A.; Anderson, J.B.; Cherukuri, P.F.; Weese-Scott, C.; Geer, L.Y.; Gwadz, M.; He, S.; Hurwitz, D.I.; Jackson, J.D.; Ke, Z.; Lanczycki, C.J.; Liebert, C.A.; Liu, C.; Lu, F.; Marchler, G.H.; Mullokandov, M.; Shoemaker, B.A.; Simonyan, V.; Song, J.S.; Thiessen, P.A.; Yamashita, R.A.; Yin, J.J.; Zhang, D. & Bryant, S.H. (2005). CDD: a conserved domain database for protein classification. *Nucl Acids Res,* Vol. 33, pp. D192–196, ISSN 0305-1048

Masashi, Y.; Watanobe, H. & Terano, A. (2001). Central regulation of hepatic function by neuropeptides. *J Gastroenterol*, Vol. 36, pp. 361–367, ISSN 0002-9270

Menschaert, G.; Vandekerckhove, T.T.; Baggerman, G.; Schoofs, L.; Luyten, W. & Van Criekinge, W. (2010). Peptidomics coming of age: a review of contributions from a bioinformatics angle. *J Proteome Res,* Vol. 9, No. 5, pp. 2051-2061, ISSN 1535-3893

Rouille, Y.; Duguay, S.J.; Lund, K.; Furuta, M.; Gong, Q.; Lipkind, G.; Oliva AA, J., Chan, S.J. & Steiner, D.F. (1995). Proteolytic processing mechanisms in the biosynthesis of neuroendocrine peptides: the subtilisin-like proprotein convertases. *Front Neuroendocrinol,* Vol. 16, No. 4, pp. 322–361, ISSN 0091-3022

Schlesinger, D.H.; Pickart, L. & Thaler, M.M. (1977). Growth-modulating serum tripeptide is glycyl-histidyl-lysine. *Experientia,* Vol. 33, No. 3, pp. 324–325, ISSN 0014-4754

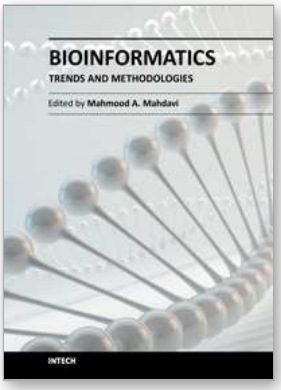
Schoofs, L. & Baggerman, G. (2003). Peptidomics in Drosophila melanogaster. *Brief Funct Genomic Proteomic,* Vol. 2, No. 2, pp. 114-120, ISSN 2041-2647

Taylor, W.R. & Jonassen, I. (2004). A structural pattern-based method for protein fold recognition. *Proteins*, Vol. 56, No. 2, pp. 222–234, ISSN 0887-3585

Vandenborne, K.; Roelens, S.A.; Darras, V.M.; Kuhn, E.R. & Van der, G.S. (2005). Cloning and hypothalamic distribution of the chicken thyrotropin-releasing hormone precursor cDNA. *J Endocrinol*, Vol. 186, No. 2, pp. 387–396, ISSN 0022-0795

Smith, R.F. & Smith, T.F. (1990). Automatic generation of primary sequence patterns from sets of related protein sequences. *Proc Natl Acad Sci USA*, Vol. 87, No. 1, pp. 118–122, ISSN 1091-6490

Ubuka, T.; Morgan, K.; Pawson, A.J.; Osugi, T.; Chowdhury, V.S.; Minakata, H.; Tsutsui, K.; Millar, R.P. & Bentley, G.E. (2009). Identification of human GnIH homologs, RFRP-1 and RFRP-3, and the cognate receptor, GPR147 in the human hypothalamic pituitary axis. *PLoS ONE*, Vol. 4, No. 12, e8400, ISSN 1932-6203



Bioinformatics - Trends and Methodologies
Edited by Dr. Mahmood A. Mahdavi

ISBN 978-953-307-282-1
Hard cover, 722 pages
Publisher InTech
Published online 02, November, 2011
Published in print edition November, 2011

Bioinformatics - Trends and Methodologies is a collection of different views on most recent topics and basic concepts in bioinformatics. This book suits young researchers who seek basic fundamentals of bioinformatic skills such as data mining, data integration, sequence analysis and gene expression analysis as well as scientists who are interested in current research in computational biology and bioinformatics including next generation sequencing, transcriptional analysis and drug design. Because of the rapid development of new technologies in molecular biology, new bioinformatic techniques emerge accordingly to keep the pace of in silico development of life science. This book focuses partly on such new techniques and their applications in biomedical science. These techniques maybe useful in identification of some diseases and cellular disorders and narrow down the number of experiments required for medical diagnostic.

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821