Online Programs and Databases of Peptides and Proteolytic Enzymes – A Brief Update for 2007–2008

Piotr Minkiewicz, Jerzy Dziuba*, Małgorzata Darewicz, Anna Iwaniak and Justyna Michalska

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

Received: April 21, 2009
Accepted: June 17, 2009

Summary

Bioinformatics methods have become one of the most important tools in peptide science. The number of available peptide databases is growing rapidly. The number of online programs able to process peptide sequences to extract information concerning their structure, physicochemical and biological properties is also increasing. Many of such programs were designed to manipulate protein sequences, but they have no built-in restrictions disabling their application to process oligopeptides containing less than 20 amino acid residues. Publications addressing these programs cannot be found in literature databases using the keyword ‘peptide’ or ‘peptides’, in connection with the term ‘bioinformatics’ or related terms, thus a reference source summarizing data from such publications seems necessary. This paper provides a brief review of bioactive peptide databases and sequence alignment programs enabling the search for peptide motifs, determination of physicochemical properties of amino acid residues, prediction of peptide structure, the occurrence of posttranslational glycosylation and immunogenicity, as well as the support of peptide design process. The review also includes databases and programs providing information about proteolytic enzymes. The databases and programs discussed in this paper were developed or updated between September 2007 and December 2008.

Key words: peptides, peptidomics, proteolysis, bioinformatics, computer databases, world wide web

Introduction

Peptide science has recently become an important area of research in chemical, biological, medical, agricultural and food sciences as well as biotechnology (1–4). The toolbox for peptide science includes bioinformatics tools (1,3,5–8). Examples of bioinformatics-aided experimental work on biologically active peptides have been described by Edwards et al. (9), who designed short peptides involved in platelet function utilizing sequence motifs associated with this activity, present in protein chains; Colette et al. (10), who investigated the occurrence of potential candidates for deorphanization of G-protein coupled receptor; Christie (11,12), who searched for precursors of peptide hormones and neuropeptides among insect protein sequences; and Southey et al. (13), who investigated the possibility of neuropeptide release from insect proteins. On the other hand, significant progress is required to obtain satisfactory results in all areas of peptide and protein science via bioinformatics methods (14–16).

This review is a continuation of our previous article (5) which covered programs developed until the summer of 2007. The present paper provides information on databases and programs useful in the study of bioactive peptide sequences, developed or updated between Sep-
Motifs corresponding to a given activity, classified as sequence similarity to the active ones as well as sequence data (sequences of bioactive peptides, considered as primary peptides, novel peptide families. It focuses only on endogenous peptides of metazoan origin, extracted from the tissues of bacteriocins, antimicrobial peptides produced by Gram-positive bacteria. A characteristic property of these substances is the presence of unusual amino acids introduced via posttranslational modifications, especially lanthionine – thus the name lantibiotics (61,62). Lantibiotics are produced, for example, by intestinal microflora and play an important role in the body’s defense against microbial pathogens. Selected lantibiotics are used for food preservation (62). Apart from information about individual peptides, the above database contains search engines based on BLAST, FASTA and Smith-Waterman algorithms, and tools for calculating the composition and physicochemical properties of peptides on the basis of their sequences.

The RAPD database (25) contains information about antimicrobial peptides obtained via recombination techniques. It provides data on the host, protocol of release as well as peptide yield. This database is useful as a source of information about the production of biologically active peptides with the use of genetic engineering, which is an emerging delivery strategy for therapeutic peptides (63).

Toxic peptides are the object of research aimed at alleviating their poisoning effect on the one hand and at their potential use as drugs on the other (64,65). The ATDB database (26) contains information about animal toxins from UniProtKB (66), related toxin databases and published literature. This database was constructed using a standardized system of annotations. Toxins in the database were annotated manually by trained biologists.

Another specialized database is ConoServer (27), which contains data on conotoxins. Conotoxins (conopeptides) produced by marine snails of the genus Conus are the object of interest as potential antiepileptic, neuroprotective, cardioprotective and analgesic drugs. The state-of-the-art in research on conotoxins has recently been reviewed by Grant et al. (67) and Han et al. (68). The database contains peptide sequences and the DNA sequences of precursors as well as peptide structures. It also provides the possibility of constructing multiple alignments and phylogenetic trees as well as of proteolysis simulation.

Bacteria and fungi synthesize numerous peptides via non-ribosomal synthetases (69). Such peptides contain atypical amino acids, possess linear or cyclic structure and reveal antimicrobial activity. Many of them are applied as antibiotics (69). These peptides are listed in the NORINE database (28).

**Peptide Databases**

A new major bioactive peptide database named PeptideDB was designed in 2008 (23). Older databases contain mainly data extracted directly from publications. This database covers all biologically active peptides and their precursors of metazoan origin, extracted from the UniProt database via BLAST searching and other *in silico* methods. The database content is not limited to the sequences of bioactive peptides, considered as primary data (59), but it also covers peptides showing high sequence similarity to the active ones as well as sequence motifs corresponding to a given activity, classified as secondary data (59). This database contains, for example, novel peptide families. It focuses only on endogenous peptides of *Metazoa* species.

Apart from general peptide databases, specialized ones have also been developed, with antimicrobial peptides as the main object of interest. Databases of antimicrobial peptides and other bioinformatics tools designed for the study of their sequences have been described by Wang (60). One of them is BACTIBASE, a new database of bacteriocins, antimicrobial peptides produced by bacteria (24). Bacteriocins are ribosomally synthesized peptides produced by Gram-positive bacteria. A characteristic property of these substances is the presence of unusual amino acids introduced *via* posttranslational modifications, especially lanthionine – thus the name lantibiotics (61,62). Lantibiotics are produced, for example, by intestinal microflora and play an important role in the body’s defense against microbial pathogens. Selected lantibiotics are used for food preservation (62). Apart from information about individual peptides, the above database contains search engines based on BLAST, FASTA and Smith-Waterman algorithms, and tools for calculating the composition and physicochemical properties of peptides on the basis of their sequences.

The RAPD database (25) contains information about antimicrobial peptides obtained *via* recombination techniques. It provides data on the host, protocol of release as well as peptide yield. This database is useful as a source of information about the production of biologically active peptides with the use of genetic engineering, which is an emerging delivery strategy for therapeutic peptides (63).

Toxic peptides are the object of research aimed at alleviating their poisoning effect on the one hand and at their potential use as drugs on the other (64,65). The ATDB database (26) contains information about animal toxins from UniProtKB (66), related toxin databases and published literature. This database was constructed using a standardized system of annotations. Toxins in the database were annotated manually by trained biologists.

Another specialized database is ConoServer (27), which contains data on conotoxins. Conotoxins (conopeptides) produced by marine snails of the genus *Conus* are the object of interest as potential antiepileptic, neuroprotective, cardioprotective and analgesic drugs. The state-of-the-art in research on conotoxins has recently been reviewed by Grant et al. (67) and Han et al. (68). The database contains peptide sequences and the DNA sequences of precursors as well as peptide structures. It also provides the possibility of constructing multiple alignments and phylogenetic trees as well as of proteolysis simulation.

Bacteria and fungi synthesize numerous peptides *via* non-ribosomal synthetases (69). Such peptides contain atypical amino acids, possess linear or cyclic structure and reveal antimicrobial activity. Many of them are applied as antibiotics (69). These peptides are listed in the NORINE database (28).

**Programs for Sequence Alignments and Motif Searching**

Recent progress in multiple sequence alignments of proteins has been reviewed by Pei (70), and by Pirovano and Heringa (71). The programs described in this review allow the search for protein fragments identical to the query peptide sequence or the construction of peptide-peptide and peptide-protein sequence alignments.

The ISPIDER Central (29) is an integrated service screening proteomic data repositories for precursors of the query peptide sequence. It does not construct alignments, but searches for protein chain fragments identical to the query sequence. The program screens databases such as the PRoteomics IDEntifications database (PRIDE) (72), PepSeeker (73), PeptideAtlas (74) and the Global Proteome Machine (75). This program enables the search across different protein sequence databases using PICR software (76).

The MAFFT program (30) provides multiple sequence alignments of nucleic acid, protein and peptide sequences. The program allows to construct phyloge-
Table 1. Peptide sequence databases and programs (23–58)

<table>
<thead>
<tr>
<th>Database or program</th>
<th>Website</th>
<th>Reference</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>BACTIBASE</td>
<td><a href="http://bactibase.pfba-lab-tun.org">http://bactibase.pfba-lab-tun.org</a></td>
<td>(24)</td>
<td>Database of antibacterial peptides (bacteriocins)</td>
</tr>
<tr>
<td>ATDB</td>
<td><a href="http://protchem.hunnu.edu.cn/toxin">http://protchem.hunnu.edu.cn/toxin</a></td>
<td>(26)</td>
<td>Database of toxic proteins and peptides of animal origin</td>
</tr>
<tr>
<td>ISPIDER Central</td>
<td><a href="http://www.ispider.manchester.ac.uk/cgi-bin/ProteomicSearch.pl">http://www.ispider.manchester.ac.uk/cgi-bin/ProteomicSearch.pl</a></td>
<td>(29)</td>
<td>Program finding information about proteins – precursors of identified peptides</td>
</tr>
<tr>
<td>MAFFT</td>
<td><a href="http://www.ebi.ac.uk/Tools/mafft/index.html">http://www.ebi.ac.uk/Tools/mafft/index.html</a></td>
<td>(30)</td>
<td>Program for the construction of multiple alignments between protein or nucleic acid sequences. Accepts also peptide sequences as queries. Information about updates published in 2008</td>
</tr>
<tr>
<td>Clustal W 2.0</td>
<td><a href="http://www.ebi.ac.uk/tools/clustalw2">http://www.ebi.ac.uk/tools/clustalw2</a></td>
<td>(31)</td>
<td>Program constructing multiple alignments between protein or peptide sequences. Information about updates published in 2007</td>
</tr>
<tr>
<td>AAIndex</td>
<td><a href="http://wwwgenome.ad.jp/aaindex">http://wwwgenome.ad.jp/aaindex</a></td>
<td>(33)</td>
<td>Database of indices describing the physicochemical properties of amino acids, suitable for the prediction of peptide properties, e.g. QSAR approach. Information about updates published in 2008</td>
</tr>
<tr>
<td>PseAAC</td>
<td><a href="http://chou.med.harvard.edu/bioinf/PseAA/">http://chou.med.harvard.edu/bioinf/PseAA/</a></td>
<td>(34)</td>
<td>Program generating the so-called pseudo amino acid composition of proteins and peptides</td>
</tr>
<tr>
<td>COPid</td>
<td><a href="http://www.imtech.res.in/raghava/copid/">http://www.imtech.res.in/raghava/copid/</a></td>
<td>(35)</td>
<td>Program designed for comparing protein amino acid and dipeptide composition as well as the content of amino acid residues with different physicochemical properties. Accepts peptide sequences as queries</td>
</tr>
<tr>
<td>PEPstr</td>
<td><a href="http://www.imtech.res.in/raghava/pepstr/">http://www.imtech.res.in/raghava/pepstr/</a></td>
<td>(36)</td>
<td>Program designed for predicting the tertiary structure of small peptides</td>
</tr>
<tr>
<td>PREDICT-2ND</td>
<td><a href="http://www.soc.acsc.edu/compbio/SAM_T08/T08-query.html">http://www.soc.acsc.edu/compbio/SAM_T08/T08-query.html</a></td>
<td>(37)</td>
<td>Program designed for peptide and protein structure prediction</td>
</tr>
<tr>
<td>HELIQUEST</td>
<td><a href="http://heliquest.ipmc.cnrs.fr">http://heliquest.ipmc.cnrs.fr</a></td>
<td>(38)</td>
<td>Program highlighting the physicochemical properties of α-helical proteins and peptides, and screening protein databases (e.g. Swiss-Prot) for protein fragments with desired properties</td>
</tr>
<tr>
<td>metaPrDOS</td>
<td><a href="http://prdos.hgc.jp/meta/">http://prdos.hgc.jp/meta/</a></td>
<td>(39)</td>
<td>Program predicting the presence of disordered regions in protein and peptide sequences</td>
</tr>
<tr>
<td>OnD-CRF</td>
<td><a href="http://babel.ucmp.umu.se/ond-crf/">http://babel.ucmp.umu.se/ond-crf/</a></td>
<td>(40)</td>
<td>Program predicting the presence of disordered regions in protein and peptide sequences</td>
</tr>
<tr>
<td>CKSAAP_OGlySite</td>
<td><a href="http://bioinformatics.cau.edu.cn/zzd_lab/CKSAAP_OGlySite/">http://bioinformatics.cau.edu.cn/zzd_lab/CKSAAP_OGlySite/</a></td>
<td>(42)</td>
<td>Program predicting the presence of mucin-type O-glycosylation sites in mammalian proteins. Accepts also peptide sequences as queries</td>
</tr>
</tbody>
</table>
etic trees for longer sequences (at least 50 residues in the case of amino acid sequences). The basic concepts and algorithms used in MAFFT as well as its comparison with other sequence alignment programs can be found in articles describing later versions of MAFFT (77, 78).

Clustal W 2.0 is the most recent version of the commonly used Clustal W program (79). This program has been rewritten in the C++ language to facilitate further development of computational procedures and adaptation to the latest versions of operating systems (31).

Motifs are defined as reproducible patterns in nucleic acid or protein sequences attributed to a given biological function. The applications of sequence motifs for protein classification and function prediction as well as bioinformatics methods used for motif construction and searching have been recently reviewed by Liu et al. (80) and Bailey (81). CompariMotif (32) is a software designed to compare two sequence motifs against each other, identifying which of them has some degree of overlapping, and determining the relationships between them. It can be used to compare a list of motifs among themselves, or with a list of previously published motifs. The program accepts oligopeptide sequences as queries.

Web Resources Exploiting the Physicochemical Properties of Amino Acid Residues

The physicochemical properties of amino acids are crucial to understand the physicochemical properties of proteins and peptides, their molecular interactions and hence biological activity (82, 83).

AAindex is a database of numerical indices representing the physicochemical and biochemical properties of amino acids and pairs of amino acids. The database consists of three sections: AAindex1 containing an amino acid index of 20 numerical values, AAindex2 containing an amino acid substitution matrix and AAindex3 containing statistical protein contact potentials (33).

Table 1. – continued

<table>
<thead>
<tr>
<th>Database or program</th>
<th>Website</th>
<th>Reference</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Glycosylation</td>
<td><a href="http://comp.chem.nottingham.ac.uk/glyco/">http://comp.chem.nottingham.ac.uk/glyco/</a></td>
<td>(43)</td>
<td>Program predicting O- and N-glycosylation sites</td>
</tr>
<tr>
<td>Predictor</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NetMHC-3.0; NetMHCpan</td>
<td><a href="http://www.cbs.dtu.dk/services/NetMHC/">http://www.cbs.dtu.dk/services/NetMHC/</a></td>
<td>(44–46)</td>
<td>Programs predicting the affinity of peptides to MHC class I</td>
</tr>
<tr>
<td>KISS</td>
<td><a href="http://cbio.ensmp.fr/kiss/">http://cbio.ensmp.fr/kiss/</a></td>
<td>(47)</td>
<td>Program predicting the binding affinity of peptides to MHC alleles</td>
</tr>
<tr>
<td>IEDB-AR</td>
<td><a href="http://tools.immunocnpitope.org">http://tools.immunocnpitope.org</a></td>
<td>(48)</td>
<td>Program designed for predicting the immunological properties of proteins. Selected options accept oligopeptide sequences as queries</td>
</tr>
<tr>
<td>EpiToolKit</td>
<td><a href="http://www.epitoolkit.org/">http://www.epitoolkit.org/</a></td>
<td>(49)</td>
<td>Program predicting T cell immune responses against proteins or peptides</td>
</tr>
<tr>
<td>rMotifGen</td>
<td><a href="http://bioinformatics.louisville.edu/brg/rMotifGen/">http://bioinformatics.louisville.edu/brg/rMotifGen/</a></td>
<td>(50)</td>
<td>Program for designing random DNA, protein or peptide sequences containing characteristic motifs</td>
</tr>
<tr>
<td>GLUE-IT</td>
<td><a href="http://guinevere.otago.ac.nz/aef/STATS/index.html">http://guinevere.otago.ac.nz/aef/STATS/index.html</a></td>
<td>(51)</td>
<td>Program estimating the diversity of peptides or proteins encoded by DNA combinatorial libraries</td>
</tr>
<tr>
<td>MOSAIC</td>
<td><a href="http://hiv.lanl.gov/content/sequence/MOSAIC/">http://hiv.lanl.gov/content/sequence/MOSAIC/</a></td>
<td>(52)</td>
<td>Program for protein or peptide vaccine design</td>
</tr>
<tr>
<td>PVS</td>
<td><a href="http://imed.med.ucm.es/PVS/">http://imed.med.ucm.es/PVS/</a></td>
<td>(53)</td>
<td>Program designed to measure site variability in groups of protein sequences. Accepts also peptide sequences as queries</td>
</tr>
<tr>
<td>Metal Detector</td>
<td><a href="http://metaldetector.dsi.unifi.it">http://metaldetector.dsi.unifi.it</a></td>
<td>(54)</td>
<td>Program predicting the potential sites of metal binding</td>
</tr>
<tr>
<td>HIVcleave</td>
<td><a href="http://chou.med.harvard.edu/bioinf/HIV/">http://chou.med.harvard.edu/bioinf/HIV/</a></td>
<td>(56)</td>
<td>Program designed for predicting HIV protease cleavage sites in proteins and peptides</td>
</tr>
<tr>
<td>PepCleave II</td>
<td><a href="http://peptibase.cs.biu.ac.il/PepCleave_II/">http://peptibase.cs.biu.ac.il/PepCleave_II/</a></td>
<td>(57)</td>
<td>Program predicting peptides resulting from proteasome cleavage</td>
</tr>
<tr>
<td>ProtIdent</td>
<td><a href="http://www.csbio.sjtu.edu.cn/bioinf/Protease/">http://www.csbio.sjtu.edu.cn/bioinf/Protease/</a></td>
<td>(58)</td>
<td>Program classifying uncharacterized proteins as proteases or non-proteases and predicting possible action mechanisms based on amino acid sequences</td>
</tr>
</tbody>
</table>

*no reference available
approach or a more generalized quantitative structure-property relationship (QSPR) approach. A review of these approaches has been recently published by Zhou et al. (83).

The Peptide Property Calculator calculates molecular mass, net charge as a function of pH, and hydrophilicity in the scale of Hopp and Woods (84). The program permits such modifications as N-terminal acetylation, biotinylation and C-terminal amidation. An example of the application of this program for the interpretation of the data obtained by chromatography can be found in Pasilis et al. (85).

The PseAAC program (34) utilizes the concept of the so-called pseudo amino acid composition based on the physicochemical properties of individual amino acid residues. This concept has been described by Chou and Shen (86,87). The amino acid sequence is replaced by the physicochemical distance between pairs of amino acids. The physicochemical distance can be calculated based on hydrophobicity, polarity and residue volume.

The COPid program (35) calculates and compares the amino acid and dipeptide content of protein and peptide sequences. It enables to determine the concentrations of amino acids of various classes based on their physicochemical properties and to construct phylogenetic trees based on amino acid and dipeptide composition.

**Programs for Peptide Structure Prediction and Visualization**

Computational methods for protein structure prediction have been reviewed by Floudas (88). The same methods may be applied for peptide structure prediction. The PEPstr program (36) serves for prediction of secondary and tertiary structure (if applicable) of peptides.

The PREDICT-2ND program (37) utilizes several algorithms described by Karplus et al. (89–91), Karchin et al. (92,93), and Shackelford and Karplus (94). This program uses an artificial neural network which is applied to numerous local structure alphabets (e.g. secondary structure propensities, local burial propensities and contacts between residues). The program also allows to compare the resulting structure with Protein Data Bank (95,96) and SCOP (97) entries covering sequence alignments and providing access to target protein structures. The methods used are continuously evaluated within CASP (Critical Assessment of Protein Structure Prediction) (98).

HELIQUEST calculates the physicochemical properties and amino acid composition of a peptide or protein fragment with α-helix structure based on its sequence, and uses the results to screen databases in order to identify protein fragments possessing similar features. In addition, the mutation module allows the user to introduce changes in the sequence to create analogues with specific properties (38). The program determines such properties as hydrophobicity, hydrophobic moment (99) and amino acid composition. The program may be useful while, for example, searching for the potential precursors of helical antimicrobial peptides (100).

The importance of intrinsic disorder in protein and peptide chains has recently been reviewed by Cortese et al. (101) and Dunker et al. (102). Two programs were designed to predict intrinsic disorder in proteins and peptides: metaPrDOS (39) and OnD-CRF (40). The first one utilizes a new prediction method for disordered regions in proteins, based on the meta approach. The method predicts the disorder tendency of each residue using Support Vector Machines from the prediction results of seven independent predictors. The other program uses the so-called conditional random fields (CRF), a new method for predicting the transition between structured and mobile or disordered regions in proteins. OnD-CRF applies CRFs relying on features which are generated from the amino acid sequence and from secondary structure prediction. Both programs were evaluated using the CASP7 targets (103).

Ramachandran plot on the web (41) displays peptide or protein structure using the Ramachandran plot. The rules governing protein structure presentation proposed by Ramachandran et al. (104) are considered as one of the most important milestones in the structural biology of proteins (105). The program can utilize pdb files generated by such programs as PEPstr (36) and SCRATCH (106).

**Programs Predicting Glycosylation Sites**

Research into protein and peptide glycosylation involves the use of bioinformatics tools (107). Examples of glycosylation site predicting programs are CKSAAP_O-GlySite (42) which predicts O-glycosylation sites in mammalian proteins, and Glycosylation Predictor (43) which may serve for O- and N-glycosylation site prediction. The first one uses the composition of k-spaced amino acid pairs (CKSAAP) based on the encoding scheme. The scheme reflects the short range interactions of amino acids within a sequence or a sequence fragment applied for predicting the physicochemical properties and interactions of proteins (108,109). The program uses the support vector machine (SVM) algorithm. The Glycosylation Predictor program predicts both O- and N-glycosylation sites via the so-called ‘random forest’ algorithm (110) based on decision trees. Several decision trees are developed using random selection of inputs and random feature selection. The trees then vote on the class for a given input. The program compares the input sequence with the known sequences around known glycosylation sites. The O-glycosylation sites used are described in the O-GLYCBASE database (111).

**Programs Predicting the Immunogenic Properties of Peptides**

The application of bioinformatics methods in immunology has recently been reviewed by Tong et al. (112) and Evans (113). Computational methods are used in two main approaches: searching for vaccines against pathogens and searching for potential allergens.

The programs NetMHC-3.0 and NetMHCpan (44–46) predict affinity to major histocompatibility complex
(MHC) class I using artificial neural networks and position-specific scoring matrices (114–116).

The KISS program (47) utilizes a support vector machine algorithm that is able to learn peptide-MHC-I binding models for many alleles simultaneously. The sharing of information is controlled by a user-defined measure of similarity between alleles. This similarity can be defined by comparing key residues known to play a role in the peptide-MHC binding.

The IEDB-AR program (48) was designed for processing protein sequences. The following options accept peptide sequences as queries: binding affinity to MHC class I and II, prediction of proteasomal cleavage, prediction of epitopes from sequence, population coverage, epitope conservancy analysis and epitope cluster analysis. Other options such as the prediction of B cell epitopes from structure and homology mapping are sufficient only for work with protein sequences or structures.

The EpiToolKit program (49) utilizes five methods for the prediction of T-cell immune responses against proteins or peptides. These methods are: SYFPEITHI (117), BIMAS/HLA_BIND (118), Epidemix (49) and Hammer (119), based on position-specific scoring matrices, as well as SVMHC (120), MHCIMulti (121) and UNITope (121) based on support vector machines.

Programs Supporting the Design of Peptides and Their Combinatorial Libraries

The design of peptides with desired structure, physicochemical and biological properties has been reviewed by Fung et al. (122), Eichler (123) and Henchey et al. (124). Combinatorial libraries are useful for finding peptides interacting with a given receptor. The combinatorial approach in peptide science has recently been reviewed by Marasco et al. (125), and by Mersich and Jungbauer (126).

As regards the programs implemented recently, rMotifGen (50) provides a method for creating random DNA or amino acid sequences with a variable number of desired motifs, where the instance of each motif can be incorporated using a position-specific scoring matrix (PSSM) or by creating an instance mutated from its corresponding consensus. The program may be applied to design the combinatorial libraries of peptides created both via phage display and chemical synthesis.

The GLUE-IT program (51) provides a statistical analysis supporting the construction of peptide combinatorial libraries. It may be used to calculate the expected number of variants represented in a given library, the library size required to obtain a given fraction of the variants or the library size required to have a given probability of sampling all possible variants. The website also contains other programs supporting the design of combinatorial libraries, such as CodonCalculator and AACalculator (127,128).

The MOSAIC program (52) serves to design artificial candidate vaccine proteins or peptides, and to assess antigen potential using the coverage of fragments being proxies for potential T cell epitopes. The vaccine design tool generates protein or peptide sequences optimized for the coverage of high-frequency fragments. The coverage-assessment tools facilitate coverage comparisons for any potential antigens.

Miscellaneous Programs Applied in Peptide Science

The PVS (Protein Variability Server) program (53) is a web-based tool that uses several variability metrics to compute absolute site variability in multiple protein or peptide sequence alignments. The variability is assigned to a reference sequence (the first sequence in the alignment or a consensus sequence). The program may be used while investigating sequence-structure-function relationships or searching for epitopes in peptide or protein chains (e.g. for vaccine design).

The Metal Detector program (54) predicts sites binding metals via histidine and cysteine residues in peptides and proteins. The program uses an artificial neural network algorithm (129).

Databases and Programs for Proteolysis Prediction

Most of biologically active peptides are encrypted in the sequences of protein precursors (130,131). The enzymatic hydrolysis of proteins is the main way to obtain bioactive peptides in living organisms. This process is commonly utilized by biotechnology and food technology. Information about proteolytic enzymes liberating and/or degrading peptides is thus crucial in peptide science.

The recent MEROPS database v. 7.8 (55) provides an integrated source of information about proteolytic enzymes. Its organizational principle is a hierarchical classification of homologous sets of peptidases grouped into families and clans. The most recent database version contains text annotations for peptidase clans and low molecular mass inhibitors. It also contains information about enzyme specificity. This database enables comparisons between enzymes of various species or subspecies.

Two specialized programs, HIVcleave (56) and PepCleave II (58), have recently been designed. The first one is aimed at predicting HIV protease cleavage sites. It is a tool supporting the search for protease inhibitors which could be potential anti-HIV drugs. The algorithm used has been described by Chou (132). The other program is aimed at identifying proteasome cleavage sites. The potential area of the application of this program is the prediction of the release of T cell epitopes containing 8–10 amino acid residues.

The ProtIdent program (58) may be helpful in searching for novel proteolytic enzymes. The program allows to discriminate between proteases and non-proteases based on the amino acid sequence. If a protein is predicted to be a proteolytic enzyme, it can attribute them to one of the six types: aspartic, cysteine, glutamic, metallo-, serine and threonine proteases. The algorithm is based on similarities between the query sequence and the known proteases of various types, and on differences between this sequence and the proteins showing no proteolytic activity.
Progress in computational amino acid sequence processing encompasses advances in protein science. The results of research conducted with the use of protein sequence databases have been reviewed by Stockinger et al. (133). The key factors affecting database usability are: programmatic access, the choice of appropriate service meeting the best practice rules and postulates, as well as data integration. The development of any database may be automated by using software designed for this purpose. SciDBMaker (134) may serve as an example of such software. Another approach facilitating database creation is automation of literature data mining.

A guide for protein function prediction on the basis of their sequence and structure has recently been proposed by Punta and Ofran (135). As regards protein function, three approaches may be followed: function annotation transfer from a sequence, function annotation transfer from secondary and tertiary structure, and de novo prediction. The first approach utilizes sequence alignments. It is obvious that the insertion, deletion or substitution of single amino acid residue has a much stronger influence on the properties of amino acid chains in the case of oligopeptides than in that of polypeptides or proteins. On the other hand, proteins revealing low similarity can contain similar short fragments. The similarity between human and viral proteins at the pentapeptide composition level (136) may serve as an example of such property. Scientists working in the area of food and nutrition are particularly interested in the shortest oligopeptides containing two or three amino acid residues because they are more easily absorbed from the digestive tract into the bloodstream than longer ones (137). Such short sequences are considered as non-informative from the perspective of the evolutionary paradigm utilizing homology searching. Among the activity predictors mentioned by Punta and Ofran (135), the short sequence signatures (motifs) are applicable to both proteins and oligopeptides. Many of the bioinformatics tools presented in this article involve machine learning methods such as artificial neural networks (ANNs) and support vector machines (SVMs). These methods are recommended for peptide bioinformatics (8). Prediction and annotation methods including those intended for determining the physicochemical properties of amino acid residues are promising. The quantitative structure-property relationship (QSPR) and the quantitative structure-activity relationship (QSAR) approaches are applicable to small molecules, e.g. to peptides containing only two amino acid residues (82,138). The binding of oligopeptides to their interactors may also be predicted using molecular docking algorithms. An example of the application of these algorithms to dipeptides that are of interest to food scientists has been described by Pripp (139). Neither of the above approaches is represented by an online program. Function annotation transfer from the structure seems to have more limited application in peptide science than in protein science due to the fact that the space of possible oligopeptide structures is much narrower than the space of possible protein structures. It seems that their activity is correlated with defined secondary structure propensities of amino acid residues building the peptide chain. The third approach proposed by Punta and Ofran (135), i.e. de novo prediction, may be applied to sequences of any length.

The approaches combining a sequence, higher order structures and amino acid property information, e.g. a simplified (for instance by omitting the evolutionary aspect) version of the so-called meta-functional signatures (140), may also be used for peptide description and activity prediction. However, such approaches are new and therefore are not represented by online computational tools.

Acknowledgements

This work was supported by the Ministry of Science and Higher Education within project no. NN 3122863 339.


121. N. Pfeifer, O. Kohlbacher, Multiple instance learning allows MHC class II epitope predictions across alleles, Lecture Notes in Computer Science, 5251 (2008) 210–221.


