ISSN 1330-9862 *review*
(FTB-2277)

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 Cieszyński 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-

*Corresponding author; Phone: ++48 89 5233 715; E-mail: jerzy.dziuba@uwm.edu.pl tember 2007 and December 2008. Algorithms and programs designed for mass spectrometry are not discussed, since progress in this area has recently been reviewed by Boonen *et al.* (*1*), Liu *et al.* (*17*), Nesvizhskii *et al.* (*18*), Matthiesen and Jensen (*19*), and Veltri (*20*). The application of bioinformatics methods for predicting peptide behaviour during chromatographic separation has been reviewed by Shinoda *et al.* (*21*).

In fact, most of the programs described here were designed for processing protein sequences, but they have no built-in restrictions excluding their application to peptides with short sequences. On the other hand, these programs (except for peptide databases) usually cannot be found on the Internet *via* literature mining using the keyword 'peptide' or 'peptides' in connection with the term 'bioinformatics' or related terms.

This review focuses on computational tools that meet the requirements established previously (*5*). The programs should process sequences considered as oligopeptides, *i.e.* containing less than 20 amino acid residues (*22*). The bioinformatics tools provide free online access, and are easy to operate by non-specialists. Data concerning proteolytic enzymes are also included. All databases and programs presented in the paper are listed in Table 1 (*23–58*). Links to them are available also *via* our website *http://www.uwm.edu.pl/biochemia* (among the links from the BIOPEP database website). Recently, the BIOPEP database has been expanded with the data of allergenic proteins, including information about the structure of their epitopes and molecular markers.

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 phylogen-

Table 1. Peptide sequence databases and programs (*23–58*)

Database or program	Website	Reference	Comment
PeptideDB	*http://www.peptides.be/*	(*23*)	Major database of biologically active peptides, peptide precursors and motifs in *Metazoa*
BACTIBASE	*http://bactibase.pfba-lab-tun.org*	(*24*)	Database of antibacterial peptides (bacteriocins)
RAPD	*http://faculty.ist.unomaha.edu/chen/rapd/index.php*	(*25*)	Database of recombinant antimicrobial peptides
ATDB	*http://protchem.hunnu.edu.cn/toxin*	(*26*)	Database of toxic proteins and peptides of animal origin
ConoServer	*http://research1t.imb.uq.edu.au/conoserver/*	(*27*)	Database of biologically active peptides of snails of the genus *Conus*
NORINE	*http://bioinfo.lifl.fr/norine/*	(*28*)	Database of non-ribosomally synthesized bioactive peptides
ISPIDER Central	*http://www.ispider.manchester.ac.uk/cgi-bin/ProteomicSearch.pl*	(*29*)	Program finding information about proteins – precursors of identified peptides
MAFFT	*http://www.ebi.ac.uk/Tools/mafft/index.html* *http://toolkit.tuebingen.mpg.de/mafft* *http://align.genome.jp/mafft/*	(*30*)	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
Clustal W 2.0	*http://www.ebi.ac.uk/tools/clustalw2*	(*31*)	Program constructing multiple alignments between protein or peptide sequences. Information about updates published in 2007
CompariMotif	*http://bioinformatics.ucd.ie/shields/software/comparimotif/*	(*32*)	Program comparing sequence motifs in proteins and/or peptides
AAIndex	*http://www.genome.ad.jp/aaindex*	(*33*)	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
Peptide Property Calculator	*http://www.innovagen.se/custom-peptide-synthesis/peptide-property-calculator/peptide-property-calculator.asp*	*	Program calculating molecular mass, charge at pH=7, average hydrophilicity and the ratio of hydrophilic amino acid residues
PseAAC	*http://chou.med.harvard.edu/bioinf/PseAA/*	(*34*)	Program generating the so-called pseudo amino acid composition of proteins and peptides
COPid	*http://www.imtech.res.in/raghava/copid/*	(*35*)	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
PEPstr	*http://www.imtech.res.in/raghava/pepstr/*	(*36*)	Program designed for predicting the tertiary structure of small peptides
PREDICT-2ND	*http://www.soe.ucsc.edu/compbio/SAM_T08/T08-query.html*	(*37*)	Program designed for peptide and protein structure prediction
HELIQUEST	*http://heliquest.ipmc.cnrs.fr*	(*38*)	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
metaPrDOS	*http://prdos.hgc.jp/meta/*	(*39*)	Program predicting the presence of disordered regions in protein and peptide sequences
OnD-CRF	*http://babel.ucmp.umu.se/ond-crf/*	(*40*)	Program predicting the presence of disordered regions in protein and peptide sequences
Ramachandran plot on the web 2.0	*http://dicsoft1.physics.iisc.ernet.in/rp/index.html*	(*41*)	Program displaying protein and peptide structure. Information about updates published in 2007
CKSAAP_OGlySite	*http://bioinformatics.cau.edu.cn/zzd_lab/CKSAAP_OGlySite/*	(*42*)	Program predicting the presence of mucin-type *O*-glycosylation sites in mammalian proteins. Accepts also peptide sequences as queries

Table 1. – continued

Database or program	Website	Reference	Comment
Glycosylation Predictor	*http://comp.chem.nottingham.ac.uk/glyco/*	(*43*)	Program predicting *O*- and *N*-glycosylation sites
NetMHC-3.0; NetMHCpan	*http://www.cbs.dtu.dk/services/NetMHC/* *http://www.cbs.dtu.dk/services/NetMHCpan-1.1/*	(*44–46*)	Programs predicting the affinity of peptides to MHC class I
KISS	*http://cbio.ensmp.fr/kiss/*	(*47*)	Program predicting the binding affinity of peptides to MHC alleles
IEDB-AR	*http://tools.immuneepitope.org*	(*48*)	Program designed for predicting the immunological properties of proteins. Selected options accept oligopeptide sequences as queries
EpiToolKit	*http://www.epitoolkit.org/*	(*49*)	Program predicting T cell immune responses against proteins or peptides
rMotifGen	*http://bioinformatics.louisville.edu/brg/rMotifGen/*	(*50*)	Program for designing random DNA, protein or peptide sequences containing characteristic motifs
GLUE-IT	*http://guinevere.otago.ac.nz/aef/STATS/index.html*	(*51*)	Program estimating the diversity of peptides or proteins encoded by DNA combinatorial libraries
MOSAIC	*http://hiv.lanl.gov/content/sequence/MOSAIC/*	(*52*)	Program for protein or peptide vaccine design
PVS	*http://imed.med.ucm.es/PVS/*	(*53*)	Program designed to measure site variability in groups of protein sequences. Accepts also peptide sequences as queries
Metal Detector	*http://metaldetector.dsi.unifi.it*	(*54*)	Program predicting the potential sites of metal binding
MEROPS	*http://merops.sanger.ac.uk*	(*55*)	Database of proteolytic enzymes. Information about updates published in 2008
HIVcleave	*http://chou.med.harvard.edu/bioinf/HIV/*	(*56*)	Program designed for predicting HIV protease cleavage sites in proteins and peptides
PepCleave II	*http://peptibase.cs.biu.ac.il/PepCleave_II/*	(*57*)	Program predicting peptides resulting from proteasome cleavage
ProtIdent	*http://www.csbio.sjtu.edu.cn/bioinf/Protease/*	(*58*)	Program classifying uncharacterized proteins as proteases or non-proteases and predicting possible action mechanisms based on amino acid sequences

*no reference available

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*). The data summarized in the AAIndex may serve for computational studies of peptide properties, for example, *via* the quantitative structure-activity relationship (QSAR) 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_OGlySite (*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*), MHCIIMulti (*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 or 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.

Final Remarks

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.

References

1. K. Boonen, B. Landuyt, G. Baggerman, S. Husson, J. Huybrechts, L. Schoofs, Peptidomics: The integrated approach of MS, hyphenated techniques and bioinformatics for neuropeptide analysis, *J. Separ. Sci. 31* (2008) 427–445.

2. E. Clynen, G. Baggerman, S. Husson, B. Landuyt, L. Schoofs, Peptidomics in drug research, *Expert Opin. Drug Discov. 3* (2008) 425–440.

3. P. Minkiewicz, J. Dziuba, M. Darewicz, A. Iwaniak, M. Dziuba, D. Nałęcz, Food peptidomics, *Food Technol. Biotechnol. 46* (2008) 1–10.

4. F. Shahidi, Y. Zhong, Bioactive peptides, *J. AOAC Int. 91* (2008) 914–931.

5. P. Minkiewicz, J. Dziuba, A. Iwaniak, M. Dziuba, M. Darewicz, BIOPEP database and other programs for processing bioactive peptide sequences, *J. AOAC Int. 91* (2008) 965–980.

6. M. Darewicz, J. Dziuba, P. Minkiewicz, Celiac disease – Background, molecular, bioinformatics and analytical aspects, *Food Rev. Int. 24* (2008) 311–329.

7. E. Petsalaki, R.B. Russell, Peptide-mediated interactions in biological systems: New discoveries and applications, *Curr. Opin. Biotechnol. 19* (2008) 244–350.

8. Z.R. Yang, Peptide bioinformatics: Peptide classification using peptide machines, *Methods Mol. Biol. 458* (2008) 159–183.

9. R.J. Edwards, N. Moran, M. Devocelle, A. Kiernan, G. Meade, W. Signac, M. Foy, S.D.E. Park, E. Dunne, D. Kenny, D.C. Shields, Bioinformatic discovery of novel bioactive peptides, *Nat. Chem. Biol. 3* (2007) 108–112.

10. J. Colette, E. Avé, B. Grenier-Boley, A.S. Coquel, K. Lesellier, K. Puget, Bioinformatics-based discovery and identification of new biologically active peptides for GPCR deorphanization, *J. Peptide Sci. 13* (2007) 568–574.

11. A.E. Christie, Neuropeptide discovery in *Ixodoidea*: An *in silico* investigation using publicly accessible expressed sequence tags, *Gen. Comp. Endocrinol. 157* (2008) 174–185.

12. A.E. Christie, *In silico* analyses of peptide paracrines/hormones in *Aphidoidea*, *Gen. Comp. Endocrinol. 159* (2008) 67–79.

13. B.R. Southey, J.V. Sweedler, S.L. Rodriguez-Zas, Prediction of neuropeptide cleavage sites in insects, *Bioinformatics, 24* (2008) 815–825.

14. A. Godzik, M. Jambon, I. Friedberg, Computational protein function prediction: Are we making progress?, *Cell. Mol. Life Sci. 64* (2007) 2505–2511.

15. U. Gowthaman, J.N. Agrewala, *In silico* tools for predicting peptides binding to HLA-class II molecules: More confusion than conclusion, *J. Proteome Res. 7* (2008) 154–163.

16. Y. Kliger, E. Gofer, A. Wool, A. Toporik, A. Apatoff, M. Olshansky, Predicting proteolytic sites in extracellular proteins: Only halfway there, *Bioinformatics, 24* (2008) 1049–1055.

17. T. Liu, M.E. Belov, N. Jaitly, W.J. Qian, R.D. Smith, Accurate mass measurements in proteomics, *Chem. Rev. 107* (2007) 3621–3653.

18. A.I. Nesvizhskii, O. Vitek, R. Aebersold, Analysis and validation of proteomic data generated by tandem mass spectrometry, *Nat. Methods, 4* (2007) 787–797.

19. R. Matthiesen, O.N. Jensen, Analysis of mass spectrometry data in proteomics, *Methods Mol. Biol. 453* (2008) 105–122.

20. P. Veltri, Algorithms and tools for analysis and management of mass spectrometry data, *Brief. Bioinform. 9* (2008) 144–155.

21. K. Shinoda, M. Sugimoto, M. Tomita, Y. Ishihama, Informatics for peptide retention properties in proteomic LC-MS, *Proteomics, 8* (2008) 787–798.

22. IUPAC-IUBMB Joint Commission on Biochemical Nomenclature, Nomenclature and symbolism for amino acids and peptides: Recommendations 1983, *Biochem. J. 219* (1984) 345–373.

23. F. Liu, G. Baggerman, L. Schoofs, G. Wets, The construction of a bioactive peptide database in *Metazoa*, *J. Proteome Res. 7* (2008) 4119–4131.

24. R. Hammami, A. Zouhir, J. Ben Hamida, I. Fliss, BACTIBASE: A new web-accessible database for bacteriocin characterization, *BMC Microbiol. 7* (2007) Article No. 89.

25. Y. Li, Z. Chen, RAPD: A database of recombinantly-produced antimicrobial peptides, *FEMS Microbiol. Lett. 289* (2008) 126–129.

26. Q.Y. He, Q.Z. He, X.C. Deng, L. Yao, E. Meng, Z.H. Liu, S.P. Liang, ATDB: A uni-database platform for animal toxins, *Nucleic Acids Res. 36* (2008) D293–D297.

27. Q. Kaas, J.C. Westermann, R. Halai, C.K.L. Wang, D.J. Craik, ConoServer, a database for conopeptide sequences and structures, *Bioinformatics, 24* (2008) 445–446.

28. S. Caboche, M. Pupin, V. Leclère, A. Fontaine, P. Jacques, G. Kucherov, NORINE: A database of nonribosomal peptides, *Nucleic Acids Res. 36* (2008) D326–D331.

29. J.A. Siepen, K. Belhajjame, J.N. Selley, S.M. Embury, N.W. Paton, C.A. Goble, S.G. Oliver, R. Stevens, L. Zamboulis, N. Martin, A. Poulovassilis, P. Jones, R. Côté, H. Hermjakob, M.M. Pentony, D.T. Jones, C.A. Orengo, S.J. Hubbard, ISPIDER Central: An integrated database web server for proteomics, *Nucleic Acids Res. 36* (2008) W485–W490.

30. K. Katoh, H. Toh, Recent developments in the MAFFT multiple sequence alignment program, *Brief. Bioinform. 9* (2008) 286–298.

31. M.A. Larkin, G. Blackshields, N.P. Brown, R. Chenna, P.A. McGettigan, H. McWilliam, F. Valentin, I.M. Wallace, A. Wilm, R. Lopez, J.D. Thompson, T.J. Gibson, D.G. Higgins, Clustal W and Clustal X version 2.0, *Bioinformatics, 23* (2007) 2947–2948.

32. R.J. Edwards, N.E. Davey, D.C. Shields, CompariMotif: Quick and easy comparisons of sequence motifs, *Bioinformatics, 24* (2008) 1307–1309.

33. S. Kawashima, P. Pokarowski, M. Pokarowska, A. Kolinski, T. Katayama, M. Kanehisa, AAindex: Amino acid index database, progress report 2008, *Nucleic Acids Res. 36* (2008) D202–D205.

34. H.B. Shen, K.C. Chou, PseAAC: A flexible web server for generating various kinds of protein pseudo amino acid composition, *Anal. Biochem. 373* (2008) 386–388.

35. M. Kumar, V. Thakur, G.P.S. Raghava, COPid: Composition based protein identification, *In Silico Biol. 8* (2008) Article No. 0011.

36. H. Kaur, A. Garg, G.P.S. Raghava, PEPstr: A *de novo* method for tertiary structure prediction of small bioactive peptides, *Protein Pept. Lett. 14* (2007) 626–631.

37. S. Katzman, C. Barrett, G. Thiltgen, R. Karchin, K. Karplus, PREDICT-2ND: A tool for generalized protein local structure prediction, *Bioinformatics, 24* (2008) 2453–2459.

38. R. Gautier, D. Douguet, B. Antonny, G. Drin, HELIQUEST: A web server to screen sequences with specific α-helical properties, *Bioinformatics, 24* (2008) 2101–2102.

39. T. Ishida, K. Kinoshita, Prediction of disordered regions in proteins based on the meta approach, *Bioinformatics, 24* (2008) 1344–1348.

40. L. Wang, U.H. Sauer, OnD-CRF: Predicting order and disorder in proteins conditional random fields, *Bioinformatics, 24* (2008) 1401–1402.

41. K. Gopalakrishnan, G. Sowmiya, S.S. Sheik, K. Sekar, Ramachandran plot on the web (2.0), *Protein Pept. Lett. 14* (2007) 669–671.

42. Y.Z. Chen, Y.R. Tang, Z.Y. Sheng, Z. Zhang, Prediction of mucin-type *O*-glycosylation sites in mammalian proteins using the composition of *k*-spaced amino acid pairs, *BMC Bioinform. 9* (2008) Article No. 101.

43. S.E. Hamby, J.D. Hirst, Prediction of glycosylation sites using random forests, *BMC Bioinform. 9* (2008) Article No. 500.

44. C. Lundegaard, K. Lamberth, M. Harndahl, S. Buus, O. Lund, M. Nielsen, NetMHC-3.0: Accurate web accessible predictions of human, mouse and monkey MHC class I affinities for peptides of length 8–11, *Nucleic Acids Res. 36* (2008) W509–W512.

45. C. Lundegaard, O. Lund, M. Nielsen, Accurate approximation method for prediction of class I MHC affinities for peptides of length 8, 10 and 11 using prediction tools trained on 9mers, *Bioinformatics, 24* (2008) 1397–1398.

46. M. Nielsen, C. Lundegaard, T. Blicher, B. Peters, A. Sette, S. Justesen, S. Buus, O. Lund, Quantitative predictions of peptide binding to any HLA-DR molecule of known sequence: NetMHCIIpan, *PLoS Comput. Biol. 4* (2008) e1000107.

47. L. Jacob, J.P. Vert, Efficient peptide–MHC-I binding prediction for alleles with few known binders, *Bioinformatics, 24* (2008) 358–366.

48. Q. Zhang, P. Wang, Y. Kim, P. Haste-Andersen, J. Beaver, P.E. Bourne, H.H. Bui, S. Buus, S. Frankild, J. Greenbaum, O. Lund, C. Lundegaard, M. Nielsen, J. Ponomarenko, A. Sette, Z. Zhu, B. Peters, Immune epitope database analysis resource (IEDB-AR), *Nucleic Acids Res. 36* (2008) W513–W518.

49. M. Feldhahn, P. Thiel, M.M. Schuler, N. Hillen, S. Stevanović, H.G. Rammensee, O. Kohlbacher, EpiToolKit – A web server for computational immunomics, *Nucleic Acids Res. 36* (2008) W519–W522.

50. E.C. Rouchka, C.T. Hardin, rMotifGen: Random motif generator for DNA and protein sequences, *BMC Bioinform. 8* (2007) Article No. 292.

51. A.E. Firth, W.M. Patrick, GLUE-IT and PEDEL-AA: New programmes for analysing protein diversity in randomized libraries, *Nucleic Acids Res. 36* (2008) W281–W285.

52. J. Thurmond, H. Yoon, C. Kuiken, K. Yusim, S. Perkins, J. Theiler, T. Bhattacharya, B. Korber, W. Fischer, Web-based design and evaluation of T-cell vaccine candidates, *Bioinformatics, 24* (2008) 1639–1640.

53. M. Garcia-Boronat, C.M. Diez-Rivero, E.L. Reinherz, P.A. Reche, PVS: A web server for protein sequence variability analysis tuned to facilitate conserved epitope discovery, *Nucleic Acids Res. 36* (2008) W35–W41.

54. M. Lippi, A. Passerini, M. Punta, B. Rost, P. Frasconi, MetalDetector: A web server for predicting metal-binding sites and disulfide bridges in proteins from sequence, *Bioinformatics, 24* (2008) 2094–2095.

55. N.D. Rawlings, F.R. Morton, C.Y. Kok, J. Kong, A.J. Barrett, MEROPS: The peptidase database, *Nucleic Acids Res. 36* (2008) D320–D325.

56. H.B. Shen, K.C. Chou, HIVcleave: A web-server for predicting human immunodeficiency virus protease cleavage sites in proteins, *Anal. Biochem. 375* (2008) 388–390.

57. I. Ginodi, T. Vider-Shalit, L. Tsaban, Y. Louzoun, Precise score for the prediction of peptides cleaved by the proteasome, *Bioinformatics, 24* (2008) 477–483.

58. K.C. Chou, H.B. Shen, ProtIdent: A web server for identifying proteases and their types by fusing functional domain and sequential evolution information, *Biochem. Biophys. Res. Commun. 376* (2008) 321–325.

59. D. Jameson, K. Garwood, C. Garwood, T. Booth, P. Alper, S.G. Oliver, N.W. Paton, Data capture in bioinformatics: Requirements and experience with Pedro, *BMC Bioinform. 9* (2008) Article No. 183.

60. G. Wang, Tool developments for structure-function studies of host defense peptides, *Protein Pept. Lett. 14* (2007) 57–69.

61. J.M. Willey, W.A. van der Donk, Lantibiotics: Peptides of diverse structure and function, *Annu. Rev. Microbiol. 61* (2007) 477–501.

62. I.F. Nes, S.S. Yoon, D.B. Diep, Ribosomally synthesized antimicrobial peptides (bacteriocins) in lactic acid bacteria: A review, *Food Sci. Biotechnol. 16* (2007) 675–690.

63. D.P. McGregor, Discovering and improving novel peptide therapeutics, *Curr. Opin. Pharmacol. 8* (2008) 616–619.

64. S. Mouhat, N. Andreotti, B. Jouriou, J.M. Sabatier, Animal toxins acting on voltage-gated potassium channels, *Curr. Pharm. Des. 14* (2008) 2503–2518.

65. A.R. Humpage, Toxin types, toxicokinetics and toxicodynamics, *Adv. Exp. Med. Biol. 619* (2008) 383–415.

66. The UniProt Consortium, The Universal Protein Resource (UniProt), *Nucleic Acids Res. 36* (2008) D190–D195.

67. M.A. Grant, K. Shanmugasundaram, A.C. Rigby, Conotoxin therapeutics: A pipeline for success?, *Expert Opin. Drug Discov. 2* (2007) 453–468.

68. T.S. Han, R.W. Teichert, B.M. Olivera, G. Bulaj, Conus venoms – A rich source of peptide-based therapeutics, *Curr. Pharm. Des. 14* (2008) 2462–2479.

69. E.A. Felnagle, E.E. Jackson, Y.A. Chan, A.M. Podevels, A.D. Berti, M.D. McMahon, M.G. Thomas, Nonribosomal peptide synthetases involved in the production of medically relevant natural products, *Mol. Pharmaceut. 5* (2008) 191–211.

70. J. Pei, Multiple protein sequence alignment, *Curr. Opin. Struct. Biol. 18* (2008) 382–386.

71. W. Pirovano, J. Heringa, Multiple sequence alignment, *Methods Mol. Biol. 452* (2008) 143–161.

72. P. Jones, R.G. Côte, L. Martens, A.F. Quinn, C.F. Taylor, W. Derache, H. Hermjakob, R. Apweiler, PRIDE: A public repository of protein and peptide identifications for the proteomics community, *Nucleic Acids Res. 34* (2006) D659–D663.

73. T. McLaughlin, J.A. Siepen, J. Selley, J.A. Lynch, K.W. Leu, H. Yin, S.J. Gaskell, S.J. Hubbard, PepSeeker: A database of proteome peptide identifications for investigating fragmentation patterns, *Nucleic Acids Res. 34* (2006) D649–D654.

74. P.G.A. Pedrioli, J.K. Eng, R. Hubley, M. Vogelzang, E.W. Deutsch, B. Raught, B. Pratt, E. Nilsson, R.H. Angeletti, R. Apweiler, K. Cheung, C.E. Costello, H. Hermjakob, S. Huang, R.K. Julian, E. Kapp, M.E. McComb, S.G. Oliver, G. Omenn, N.W. Paton, R. Simpson, R. Smith, C.F. Taylor, W. Zhu, R. Aebersold, A common open representation of mass spectrometry data and its application to proteomics research, *Nat. Biotechnol. 22* (2004) 1459–1466.

75. R. Craig, J.P. Cortens, R.C. Beavis, Open source system for analyzing, validating, and storing protein identification data, *J. Proteome Res. 3* (2004) 1234–1242.

76. R.G. Côté, P. Jones, L. Martens, S. Kerrien, F. Reisinger, Q. Lin, R. Leinonen, R. Apweiler, H. Hermjakob, The Protein Identifier Cross-Referencing (PICR) service: Reconciling protein identifiers across multiple source databases, *BMC Bioinform. 8* (2007) Article No. 401.

77. K. Katoh, K. Misawa, K. Kuma, T. Miyata, MAFFT: A novel method for rapid multiple sequence alignment based on fast Fourier transform, *Nucleic Acids Res. 30* (2002) 3059–3066.

78. K. Katoh, K. Kuma, H. Toh, T. Miyata, MAFFT version 5: Improvement of accuracy of multiple sequence alignment, *Nucleic Acids Res. 33* (2005) 511–518.

79. J.D. Thompson, D.G. Higgins, T.J. Gibson, Clustal W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice, *Nucleic Acids Res. 22* (1994) 4673–4680.

80. Z.P. Liu, L.Y. Wu, Y. Wang, X.S. Zhang, L. Chen, Bridging protein local structures and protein functions, *Amino Acids, 35* (2008) 627–650.

81. T.L. Bailey, Discovering sequence motifs, *Methods Mol. Biol. 452* (2008) 231–251.

82. J.C. Biro, The proteomic code: A molecular recognition code for proteins, *Theor. Biol. Med. Model. 4* (2007) Article No. 45.

83. P. Zhou, F.F. Tian, Y.Q. Wu, Z.L. Li, Z.C. Shang, Quantitative Sequence-Activity Model (QSAM): Applying QSAR strategy to model and predict bioactivity and function of peptides, proteins and nucleic acids, *Curr. Comput. Aided Drug Des. 4* (2008) 311–321.

84. T.P. Hopp, K.R. Woods, Prediction of protein antigenic determinants from amino acid sequences, *Proc. Natl Acad. Sci. USA, 78* (1981) 3824–3828.

85. S.P. Pasilis, V. Kertesz, G.J. Van Berkel, M. Schulz, S. Schorcht, Using HPTLC/DESI-MS for peptide identification in 1D separations of tryptic protein digests, *Anal. Bioanal. Chem. 391* (2008) 317–324.

86. K.C. Chou, Prediction of protein subcellular locations by incorporating quasi-sequence-order effect, *Biochem. Biophys. Res. Commun. 278* (2000) 477–483.

87. K.C. Chou, H.B. Shen, Recent progress in protein subcellular localization prediction, *Anal. Biochem. 370* (2007) 1–16.

88. C.A. Floudas, Computational methods in protein structure prediction, *Biotechnol. Bioeng. 97* (2007) 207–213.

89. K. Karplus, R. Karchin, C. Barrett, S. Tu, M. Cline, M. Diekhans, L. Grate, J. Casper, R. Hughey, What is the value added by human intervention in protein structure prediction?, *Proteins* (Suppl. 5), *45* (2001) 86–91.

90. K. Karplus, R. Karchin, J. Draper, J. Casper, Y. Mandel-Gutfreund, M. Diekhans, R. Hughey, Combining local-structure, fold-recognition, and new-fold methods for protein structure prediction, *Proteins* (Suppl. 6), *53* (2003) 491–496.

91. K. Karplus, S. Katzman, G. Shackelford, M. Koeva, J. Draper, B. Barnes, M. Soriano, R. Hughey, SAM-T04: What is new in protein-structure prediction for CASP6, *Proteins* (Suppl. 7), *61* (2005) 135–142.

92. R. Karchin, M. Cline, Y. Mandel-Gutfreund, K. Karplus, Hidden Markov models that use predicted local structure for fold recognition: Alphabets of backbone geometry, *Proteins, 51* (2003) 504–514.

93. R. Karchin, M. Cline, K. Karplus, Evaluation of local structure alphabets based on residue burial, *Proteins, 55* (2004) 508–518.

94. G. Shackelford, K. Karplus, Contact prediction using mutual information and neural nets, *Proteins* (Suppl. 8), *69* (2007) 159–164.

95. H.M. Berman, K. Henrick, H. Nakamura, J.L. Markley, The worldwide Protein Data Bank (wwPDB): Ensuring a single, uniform archive of PDB data, *Nucleic Acids Res. 35* (2007) D301–D303.

96. K. Henrick, Z. Feng, W.F. Bluhm, D. Dimitropoulos, J.F. Doreleijers, S. Dutta, J.L. Flippen-Anderson, J. Ionides, C. Kamada, E. Krissinel, C.L. Lawson, J.L. Markley, H. Nakamura, R. Newman, Y. Shimizu, J. Swaminathan, S. Velankar, J. Ory, E.L. Ulrich, W. Vranken, J. Westbrook, R. Yamashita, H. Yang, J. Young, M. Yousufuddin, H.M. Berman, Remediation of the protein data bank archive, *Nucleic Acids Res. 36* (2008) D426–D433.

97. A. Andreeva, D. Howorth, J.M.M. Chandonia, S.E. Brenner, T.J.P. Hubbard, C. Chothia, A.G. Murzin, Data growth and its impact on the SCOP database: New developments, *Nucleic Acids Res. 36* (2008) D419–D425.

98. J. Moult, K. Fidelis, A. Kryshtafovych, B. Rost, T. Hubbard, A. Tramontano, Critical assessment of methods of protein structure prediction – Round VII, *Proteins* (Suppl. 8), *69* (2007) 3–9.

99. D. Eisenberg, R.M. Weiss, T.C. Terwilliger, The helical hydrophobic moment: A measure of the amphiphilicity of a helix, *Nature, 299* (1982) 371–374.

100. M. Dathe, T. Wieprecht, Structural features of helical antimicrobial peptides: Their potential to modulate activity on model membranes and biological cells, *Biochim. Biophys. Acta, 1462* (1999) 71–87.

101. M.S. Cortese, V.N. Uversky, A.K. Dunker, Intrinsic disorder in scaffold proteins: Getting more from less, *Progr. Biophys. Mol. Biol. 98* (2008) 85–106.

102. A.K. Dunker, I. Silman, V.N. Uversky, J.L. Sussman, Function and structure of inherently disordered proteins, *Curr. Opin. Struct. Biol. 18* (2008) 756–764.

103. L. Bordoli, F. Kiefer, T. Schwede, Assessment of disorder predictions in CASP7, *Proteins* (Suppl. 8), *69* (2007) 129–136.

104. G.N. Ramachandran, C. Ramakrishnan, V. Sasisekharan, Stereochemistry of polypeptide chain configurations, *J. Mol. Biol. 7* (1963) 95–99.

105. A.N. Lupas, The long coming of computational structural biology, *J. Struct. Biol. 163* (2008) 254–257.

106. J. Cheng, A.Z. Randall, M.J. Sweredowski, P. Baldi, SCRATCH: A protein structure and structural feature prediction server, *Nucleic Acids Res. 33* (2005) W72–W77.

107. D. Wang, S. Gedela, Insights of new tools in glycomic research, *J. Proteomics Bioinform. 1* (2008) 374–378.

108. K. Chen, L. Kurgan, M. Rahbari, Prediction of protein crystallization using collocation of amino acid pairs, *Biochem. Biophys. Res. Commun. 355* (2007) 764–769.

109. K. Chen, L.A. Kurgan, J. Ruan, Prediction of flexible/rigid regions from protein sequences using *k*-spaced amino acid pairs, *BMC Struct. Biol. 7* (2007) Article No. 25.

110. L. Breiman, Random forests, *Machine Learning, 45* (2001) 5–32.

111. R. Gupta, H. Birch, K. Rapacki, S. Brunak, J.E. Hansen, O-GLYCBASE version 4.0: A revised database of *O*-glycosylated proteins, *Nucleic Acids Res. 27* (1999) 370–372.

112. J.C. Tong, T.W. Tan, S. Ranganathan, Methods and protocols for prediction of immunogenic epitopes, *Brief. Bioinform. 8* (2007) 96–108.

113. M.C. Evans, Recent advances in immunoinformatics: Application of *in silico* tools to drug development, *Curr. Opin. Drug Discov. Develop. 11* (2008) 233–241.

114. S. Buus, S.L. Lauemoller, P. Worning, C. Kesmir, T. Frimurer, S. Corbet, A. Fomsgaard, J. Hilden, A. Holm, S. Brunak, Sensitive quantitative predictions of peptide-MHC binding by a 'Query by Committee' artificial neural network approach, *Tissue Antigens, 62* (2003) 378–384.

115. M. Nielsen, C. Lundegaard, P. Worning, S.L. Lauemøller, K. Lamberth, S. Buus, S. Brunak, O. Lund, Reliable prediction of T-cell epitopes using neural networks with novel sequence representations, *Protein Sci. 12* (2003) 1007–1017.

116. M. Nielsen, C. Lundegaard, P. Worning, C.S. Hvid, K. Lamberth, S. Buus, S. Brunak, O. Lund, Improved prediction of MHC class I and class II epitopes using a novel Gibbs sampling approach, *Bioinformatics, 20* (2004) 1388–1397.

117. H.G. Rammensee, J. Bachmann, N.P.N. Emmerich, O.A. Bachor, S. Stevanović, SYFPEITHI: Database for MHC ligands and peptide motifs, *Immunogenetics, 50* (1999) 213–219.

118. K.C. Parker, M.A. Bednarek, J.E. Coligan, Scheme for ranking potential HLA-A2 binding peptides based on independent binding of individual peptide side-chains, *J. Immunol. 152* (1994) 163–175.

119. T. Sturniolo, E. Bono, J. Ding, L. Raddrizzani, O. Tuereci, U. Sahin, M. Braxenthaler, F. Gallazzi, M.P. Protti, F. Sinigaglia, J. Hammer, Generation of tissue-specific and promiscuous HLA ligand databases using DNA microarrays and virtual HLA class II matrices, *Nat. Biotechnol. 17* (1999) 555–561.

120. P. Dönnes, O. Kohlbacher, SVMHC: A server for prediction of MHC-binding peptides, *Nucleic Acids Res. 34* (2006) W194–W197.

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.

122. H.K. Fung, W.J. Welsh, C.A. Floudas, Computational *de novo* peptide and protein design: Rigid templates versus flexible templates, *Ind. Eng. Chem. Res. 47* (2008) 993–1001.

123. J. Eichler, Peptides as protein binding site mimetics, *Curr. Opin. Chem. Biol. 12* (2008) 707–713.

124. L.K. Henchey, A.L. Jochim, P.S. Arora, Contemporary strategies for the stabilization of peptides in the α-helical conformation, *Curr. Opin. Chem. Biol. 12* (2008) 692–697.

125. D. Marasco, G. Perretta, M. Sabatella, M. Ruvo, Past and future perspectives of synthetic peptide libraries, *Curr. Protein Pept. Sci. 9* (2008) 447–467.

126. C. Mersich, A. Jungbauer, Generation of bioactive peptides by biological libraries, *J. Chromatogr. B, 861* (2008) 160–170.

127. A.E. Firth, W.M. Patrick, Statistics of protein library construction, *Bioinformatics, 21* (2005) 3314–3315.

128. W.M. Patrick, A.E. Firth, Strategies and computational tools for improving randomized protein libraries, *Biomol. Eng. 22* (2005) 105–112.

129. A. Passerini, M. Punta, A. Ceroni, B. Rost, P. Frasconi, Identifying cysteines and histidines in transition metal binding sites using support vector machines and neural networks, *Proteins, 65* (2006) 305–316.

130. E. Schlimme, H. Meisel, Bioactive peptides derived from milk proteins. Structural, physiological and analytical aspects, *Nahrung, 39* (1995) 1–20.

131. A.A. Karelin, E.Y. Blischenko, V.T. Ivanov, A novel system of peptidergic regulation, *FEBS Lett. 428* (1998) 7–12.

132. K.C. Chou, Prediction of HIV protease cleavage sites in proteins, *Anal. Biochem. 233* (1996) 1–14.

133. H. Stockinger, T. Attwood, S.N. Chohan, R. Côté, P. Cudré-Mauroux, L. Falquet, P. Fernandes, R.D. Finn, T. Hupponen, E. Korpelainen, A. Labarga, A. Laugraud, T. Lima, E.

Pafilis, M. Pagni, S. Pettifer, I. Phan, N. Rahman, Experience using web services for biological sequence analysis, *Brief. Bioinform. 9* (2008) 493–505.

134. R. Hammami, A. Zouhir, K. Nagmouchi, J. Ben Hamida, I. Fliss, SciDBMaker: New software for computer-aided design of specialized biological databases, *BMC Bioinform. 9* (2008) Article No. 121.

135. M. Punta, Y. Ofran, The rough guide to *in silico* function prediction, or how to use sequence and structure information to predict protein function, *PLoS Comput. Biol. 4* (2008) e1000160.

136. D. Kanduc, A. Stufano, G. Lucchese, A. Kusalik, Massive peptide sharing between viral and human proteomes, *Peptides, 29* (2008) 1755–1766.

137. V. Vermeirssen, J. Van Camp, W. Verstraete, Bioavailability of angiotensin I converting enzyme inhibitory peptides, *Br. J. Nutr. 92* (2004) 357–366.

138. Q.S. Du, R.B. Huang, K.C. Chou, Recent advances in QSAR and their applications in predicting the activities of chemical molecules, peptides and proteins for drug design, *Curr. Protein Pept. Sci. 9* (2008) 248–260.

139. A.H. Pripp, Docking and virtual screening of ACE inhibitory dipeptides, *Eur. Food Res. Technol. 225* (2007) 589–592.

140. K. Wang, J.A. Horst, G. Cheng, D.C. Nickle, R. Samudrala, Protein meta-functional signatures from combining sequence, structure, evolution and amino acid property information, *PLoS Comput. Biol. 4* (2008) e1000181.