

Searching for hypothetical proteins: Theory and practice based upon original data and literature

Gert Lubec^{*}, Leila Afjehi-Sadat, Jae-Won Yang, Julius Paul Pradeep John

Department of Pediatrics, Division of Basic Sciences, Medical University of Vienna, Waehringer Guertel 18-20, A-1090, Vienna, Austria

Received 23 May 2005; received in revised form 18 September 2005; accepted 2 October 2005

Abstract

A large part of mammalian proteomes is represented by hypothetical proteins (HP), i.e. proteins predicted from nucleic acid sequences only and protein sequences with unknown function. Databases are far from being complete and errors are expected.

The legion of HP is awaiting experiments to show their existence at the protein level and subsequent bioinformatic handling in order to assign proteins a tentative function is mandatory. Two-dimensional gel-electrophoresis with subsequent mass spectrometrical identification of protein spots is an appropriate tool to search for HP in the high-throughput mode. Spots are identified by MS or by MS/MS measurements (MALDI-TOF, MALDI-TOF-TOF) and subsequent software as e.g. Mascot or ProFound. In many cases proteins can thus be unambiguously identified and characterised; if this is not the case, de novo sequencing or Q-TOF analysis is warranted. If the protein is not identified, the sequence is being sent to databases for BLAST searches to determine identities/similarities or homologies to known proteins. If no significant identity to known structures is observed, the protein sequence is examined for the presence of functional domains (databases PROSITE, PRINTS, InterPro, ProDom, Pfam and SMART), subjected to searches for motifs (ELM) and finally protein–protein interaction databases (InterWeaver, STRING) are consulted or predictions from conformations are performed.

We here provide information about hypothetical proteins in terms of protein chemical analysis, independent of antibody availability and specificity and bioinformatic handling to contribute to the extension/completion of protein databases and include original work on HP in the brain to illustrate the processes of HP identification and functional assignment.

© 2005 Elsevier Ltd. All rights reserved.

Keywords: Hypothetical protein; Bioinformatics; Prediction; Putative function; Tentative structure

Contents

1. What are hypothetical proteins?	91
1.1. Definitions	91
1.2. Glossary	92

Abbreviations: 2-DE, two-dimensional gel electrophoresis; 3D-PSSM, three-dimensional position-specific scoring matrix; ANN, artificial neural networks; BIND, biomolecular interaction network database; BLAST, basic local alignment search tool; CAMPASS, Cambridge database of protein alignments organised as structural superfamilies; CATH, protein structure classification; CDD, conserved domain database; CHAPS, 3-[(3-cholamidopropyl)dimethylammonio]-1-propane-sulfonate; COG, clusters of orthologous groups of proteins; DIP, database of interacting proteins; EDTA, ethylenediaminetetraacetic acid; ELM, eukaryotic linear motif server; FSSP, database of families of structurally similar proteins; GRAVY, grand average of hydropathy; HOMSTRAD, homologous structure alignment database; HP, hypothetical protein; LC, liquid chromatography; LRR, leucine rich repeats; MALDI, matrix-assisted laser desorption/ionisation; MB, medulloblastoma; MS, mass spectrometry; MS/MS, tandem mass spectrometry; NMR, nuclear magnetic resonance; NNA, neural network algorithm; OGP, octyl β -D-glucopyranoside; ORF, open reading frames; ORFans, orphan ORFs; PBS, phosphate buffered saline; Pfam, protein family databases of alignments and HMMs; PHI-BLAST, pattern hit initiated BLAST; PIRSF, protein information resource superfamily; PMF, peptide mass fingerprint; PSI-BLAST, position specific iterative BLAST; Q-TOF, quadrupole time-of-flight; RPS-BLAST, reversed position specific BLAST; SCOP, structural classification of proteins; SLE, supervised locally linear embedding; SMART, simple modular architecture research tool; SP, signal peptide; SVM, support vector machine; TFA, trifluoroacetic acid; TOF, time of flight; UPF, uncharacterised protein family

^{*} Corresponding author. Tel.: +43 1 40400 3215; fax: +43 1 40400 3194.

E-mail address: gert.lubec@meduniwien.ac.at (G. Lubec).

1.3.	Relevance and importance of hypothetical proteins	92
2.	Methodology for hunting hypothetical proteins	92
2.1.	Cell culture, sample preparation and two-dimensional gel electrophoresis	92
2.1.1.	Cell culture and sample preparation	93
2.1.2.	Two-dimensional gel electrophoresis (2-DE)	93
2.2.	Mass spectrometrical protein chemical identification.	93
2.2.1.	In gel digestion	93
2.2.2.	MALDI-TOF and MALDI-TOF/TOF-mass spectrometry (MS)	94
2.3.	Results from the chemical analysis of HPs	94
3.	Data mining	94
3.1.	Methods for protein identification from mass spectrometrical data/spectra	94
3.1.1.	Mascot searches	96
3.1.2.	Protein Prospector searches	96
3.1.3.	ProFound searches	98
3.2.	Alignments and BLAST searches	98
3.2.1.	Alignments	98
3.2.2.	BLAST searches	99
4.	Functional analysis of hypothetical proteins	105
4.1.	Domain analysis–databases	105
4.2.	Motif analysis	106
4.3.	Functional analysis by protein–protein interaction and protein association databases.	109
5.	Homology searches	109
6.	Protein characterisation by physicochemical properties	112
6.1.	Amino acid composition and protein stability	112
6.2.	Hydrophobicity	112
6.2.1.	ProtParam tool	118
7.	Prediction of subcellular localisation	118
8.	Signal peptide prediction	118
9.	Membrane proteins prediction	119
9.1.	TMHMM, HMMTOP and PHD programs	119
10.	Structural bioinformatics.	119
10.1.	Homology modeling	120
10.1.1.	SWISS-MODEL server	120
10.2.	Threading methods	121
10.3.	Hybrid methods	121
10.4.	Practical ab initio methods	123
11.	Conclusion	123
	Acknowledgements	123
	Appendix	123
	References	123

1. What are hypothetical proteins?

1.1. Definitions

There is so far no classification of hypothetical proteins (HPs) and working terms are replacing definitions of hypothetical proteins. In the strict sense, HPs are predicted proteins, proteins predicted from nucleic acid sequences and that have not been shown to exist by experimental protein chemical evidence. Moreover, these proteins are characterised by low identity to known, annotated proteins. In an attempt to define HPs Galperin (2001) and Galperin and Koonin (2004) defined “*conserved hypothetical proteins*” as a large fraction of genes in sequenced genomes encoding those that are found in organisms from several phylogenetic lineages but have not been functionally characterised and described at the protein chemical level. These structures may represent up to half of the potential protein coding regions of a genome.

Shmueli et al. (2004) mention a representative fraction, however, consisting of HPs lacking any significant sequence similarity to other ORFs in the databases and are termed orphan ORFs (syn: ORFans) or “*poorly conserved ORFs*” (Siew and Fischer, 2003a,b). And it is these HP that form a myriad of structures that match no other sequence in the databases. Over half of the ORFans are shorter than 50 amino acid residues and the probability that these HPs are expressed is rather low (Siew and Fischer, 2003a,b).

Another possibility for HP classification would be based upon the presence or absence of a known gene name for the HP. In Table 5, we present three HP where no gene name could be detected in public available databases. While the protein name is provided for putative 55 kDa protein, no gene name is available and the annotation was based upon a nucleotide sequence obtained from adrenal gland by direct submission (Gu et al., 1999: EMBL/GenBank/DBJ; <http://www.ncbi.nlm.nih.gov/>). We here propose to consider

this system as one non-sophisticated possible way to classify HPs.

1.2. Glossary

Apart from the absence of comprehensive definitions of HPs there is a need to provide a glossary as can be seen in literature and this fact causes major confusions and misunderstandings in proteomic data mining and data handling. Moreover, understanding terminology is a prerequisite for coping with ontologies in bioinformatics (Stevens et al., 2000).

Even basic terms as identity, similarity and homology are misused and are causing misinterpretations and generating a “terminology muddle” (Reeck et al., 1987). Identity (sequence identity) is the extent to which two amino acid (or nucleotide) sequences are invariant. Percentage identity is a frequently quoted statistics for an alignment of two sequences (Rost, 1999).

Similarities are sequence similarities or other types of similarities and are the result of statistical analysis describing a level or degree of similarity, an alignment with optimized similarity or the percentage of positional identity in an alignment as well as evaluating the probability associated with an alignment. Similar pairs of residues are structurally or functionally related including conservative substitutions. Amino acids with similar properties include the basic amino acids (K, R, H), acidic amino acids (D, E), hydrophobic amino acids (W, F, Y, L, I, V, M, A), etc. Percentage similarity of two protein sequences is the sum of both identical and similar matches. This may indicate that identity is a more valuable and reliable description and therefore should be used to determine relationships between sequences.

Two proteins are said to be homologous if they possess a common evolutionary origin (Fitch, 1970). Evidence for homology should be explicitly laid out, making it clear that the proposed relationship is based on the level of observed similarity, the statistical significance of the similarity and possibly other lines of reasoning (Reeck et al., 1987).

Orthologs arise from speciation, the same gene in different organisms. Paralogs may be derived from duplication within a genome.

Proteins that have no common ancestors but possess structural similarity are called analogs (Dokholyan and Shakhnovich, 2001).

Equivalogs are proteins with equivalent functions.

A motif (protein sequence motif) is a set of conserved amino acid residues that are important for function.

A domain is a structurally compact, independently folding unit forming a stable three-dimensional structure. Typically, a conserved domain contains one or more motifs (Koonin and Galperin, 2003; <http://www.ncbi.nlm.nih.gov/books/bv.fcgi?rid=sef.section.70>). Another misuse of wording has to be clarified: a protein exhibiting extensive sequence similarity to a characterised protein and/or the same conserved regions is labeled “probable”. “Putative” is used for proteins that exhibit

limited sequence similarity to characterised proteins—these proteins often have a conserved site but no other significant similarity to a characterised protein. The label “potential” is used to indicate assignment by comparative analysis and/or prediction programs for signal sequences, transmembrane regions, coiled-coils, etc. without experimental proof (Junker et al., 1999).

Unknown proteins are structures that have been experimentally shown to exist but are not characterised in protein chemical terms or cannot be linked to a known gene. Among these are protein classes that present with specific domains as e.g. uncharacterised protein families (UPFs).

Another protein category includes proteins of unknown functions. These are experimentally documented but no known functional or structural domain is observed. They may contain sequences including coiled-coil structures or transmembrane regions that do not allow assignment of function.

1.3. Relevance and importance of hypothetical proteins

About half the proteins in most genomes are candidates for HPs (Minion et al., 2004). This group is of utmost importance to complete genomic and proteomic information. Detection of new HPs not only offers presentation of new structures but also new functions. There will be new structures with so far unknown conformations and new domains and motifs will be arising. A series of additional protein pathways and cascades will be revealed, completing our fragmentary knowledge on the mosaic of proteins per se. The network of protein–protein interactions will be increasing logarithmically. New HPs may be serving as markers and pharmacological targets. Last not least, detection of HP would be of benefit to genomics enabling the discovery of so far unknown or even predicted genes.

2. Methodology for hunting hypothetical proteins

The heart of the searching strategy is to separate a HP amino acid sequence as long (complete) as possible without modifications or degradations prior to the analytical process. Therefore, methods degrading or enzymatic cleavage of a HP have to be avoided and therefore a technique using trypsin cleavage of a protein mixture—as already proposed—rather than separating a protein that is subsequently subject to proteolysis to generate peptides for mass spectrometrical analysis, is inappropriate. A HP could never be identified based upon a single peptide.

2.1. Cell culture, sample preparation and two-dimensional gel electrophoresis

Fair separation of a protein mixture forms the basis for mass spectrometrical analysis of a HP. Following prefractionation by chromatographical steps or differential centrifugations as e.g. in the case of subcellular fractionation we propose the technique of two-dimensional gel electrophoresis with Coomassie staining as a suitable method for protein hunting (Lubec

et al., 2003; Afjehi-Sadat et al., 2004, 2005; Oh et al., 2004; Shin et al., 2004a,b,c).

Coomassie staining only reveals high abundance proteins when no prefractionation is used but it reliably enables mass spectrometrical analysis by providing sufficient protein to cope with the relatively low sensitivity (but high specificity) of most MS techniques.

Using this technique high protein identification rates can be reached.

The use of more sensitive stains is not recommended as this only leads to detection of more spots but as the protein amount in (silver stained, fluorescent dyes) visible spots does not reach threshold levels for identification of proteins, this approach may be redundant. In addition, some stains are leading to miscleavages and thus the identification rate is reduced.

2.1.1. Cell culture and sample preparation

The DAOY cell line (ATCC: HTB-186; Jacobsen et al., 1985) was cultivated according to specific ATCC guidelines (<http://www.lgcpromochem-atcc.com/SearchCatalogs/lor>). Harvested cells were washed three times with 10 mL PBS (phosphate buffered saline) (Gibco BRL, Gaithersburg, MD, USA) and centrifuged for 10 min at $800 \times g$ at room temperature. The supernatant was discarded and the pellet was suspended in 1.0 ml of sample buffer consisting of 40 mM Tris, 7 M urea (Merck, Darmstadt, Germany), 2 M thiourea (Sigma, St. Louis, MO, USA), 4% CHAPS (3-[(3-cholamidopropyl)dimethylammonio]-1-propane-sulfonate) (Sigma, St. Louis, MO, USA), 65 mM 1,4-dithioerythritol (Merck, Germany), 1 mM EDTA (ethylenediaminetetraacetic acid) (Merck, Germany), protease inhibitors complete (Roche, Basel, Switzerland) and 1 mM phenylmethylsulfonyl chloride. The suspension was sonicated for approximately 30 s. After homogenisation samples were left at room temperature for 1 h and centrifuged at 14,000 rpm for 1 h. The supernatant was transferred into Ultrafree-4 centrifugal filter unit (Millipore, Bedford, MA), for desalting and concentrating proteins. Protein content of the supernatant was quantified by Bradford protein assay system (Bradford, 1976). The standard curve was generated using bovine serum albumin and absorbance was measured at 595 nm.

2.1.2. Two-dimensional gel electrophoresis (2-DE)

2-DE was performed essentially as reported (Weitzdoerfer et al., 2002; Yang et al., 2004). Samples of 550 μ g protein were applied on immobilized *pI* 3–10 nonlinear gradient strips (Amersham Bioscience, Uppsala, Sweden). Focusing started at 200 V and the voltage was gradually increased to 8000 V at 4 V/min and kept constant for a further 3 h (approximately 150,000 V h totally). The second-dimensional separation was performed on 9–16% gradient sodium SDS polyacrylamide gels. After protein fixation for 12 h in 50% methanol and 10% acetic acid, the gels were stained with colloidal Coomassie blue (Novex, San Diego, CA, USA) for 8 h and excess of dye was washed out from the gels with distilled water. Molecular masses were determined by running standard protein markers (Bio-Rad Laboratories, Hercules, CA, USA), covering the range 10–

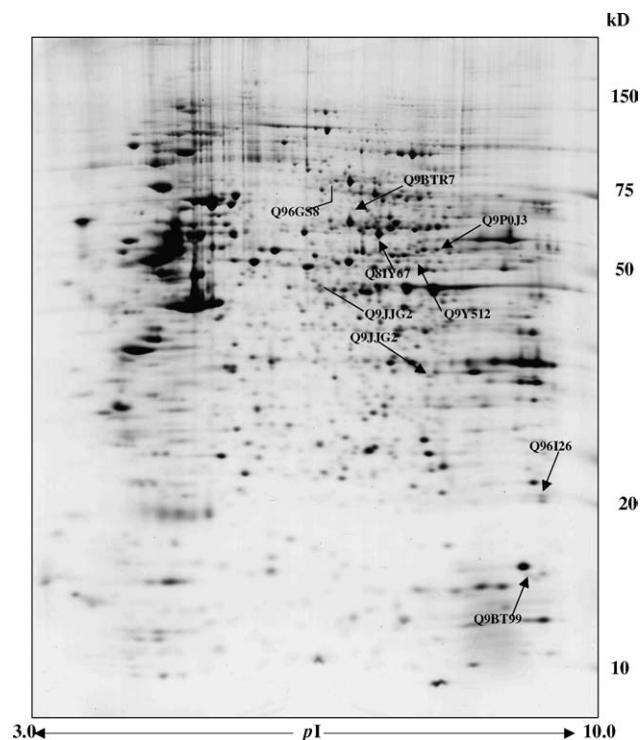


Fig. 1. Map of eight hypothetical proteins observed in the DAOY cell line. Accession numbers (Swissprot) are provided. Protein Q9JIG2 is represented by two spots, probably representing splicing variants or posttranslational modifications.

250 kDa. *pI* values were used as given by the supplier of the immobilized *pI* gradient strips.

Fig. 1 shows HPs of a partial proteome from a medulloblastoma cell line that were unambiguously identified by MS and MS/MS.

2.2. Mass spectrometrical protein chemical identification

There are three main principle methods that are suitable; MALDI-MS and MALDI-MS/MS allow high-throughput analysis of two-dimensional gels (384 protein spots can be analysed from one target) while Q-TOF and LC-MS (-MS) currently cannot be used for this purpose and do not represent strategies in the search for HPs.

2.2.1. In gel digestion

Spots were excised with a spot picker (PROTEINEER spTM, Bruker Daltonics, Leipzig, Germany), placed into 96-well microtiter plates and in-gel digestion and sample preparation for MALDI analysis were performed by an automated procedure (PROTEINEER dpTM, Bruker Daltonics) (Suckau et al., 2003; Yang et al., 2004). Briefly, spots were excised and washed with 10 mM ammonium bicarbonate and 50% acetonitrile in 10 mM ammonium bicarbonate. After washing, gel plugs were shrunk by addition of acetonitrile and dried by blowing out the liquid through the pierced well bottom. The dried gel pieces were reswollen with 40 ng/ μ l trypsin (Promega, Madison, WI, USA) in digestion

buffer (consisting of 5 mM Octyl β -D-glucopyranoside (OGP) and 10 mM ammonium bicarbonate) and incubated for 4 h at 30 °C. In order to improve the sequence coverage, HP Q9Y512 spots were further digested with Lys-C and Asp-N. Lys-C digests were performed by addition of 5 mM ammonium bicarbonate containing 30 ng/ μ l of Lys-C (sequencing grade; Roche Diagnostic, Mannheim, Germany). Asp-N digestions were performed by addition of 25 mM ammonium bicarbonate containing 12.5 ng/ μ l of Asp-N (sequencing grade; Roche Diagnostic, Mannheim, Germany) and both the digests incubated for 18 h at 37 °C. Extraction was performed with 10 μ l of 1% TFA in 5 mM OGP.

2.2.2. MALDI-TOF and MALDI-TOF/TOF-mass spectrometry (MS)

Extracted peptides were directly applied onto a target (AnchorChipTM, Bruker Daltonics) that was loaded with α -cyano-4-hydroxy-cinnamic acid (Bruker Daltonics) matrix thinlayer. The mass spectrometer used in this work was an UltraflexTM TOF/TOF (Bruker Daltonics) operated in the reflector for MALDI-TOF peptide mass fingerprint (PMF) or LIFT mode for MALDI-TOF/TOF with a fully automated mode using the FlexControlTM software. An accelerating voltage of 25 kV was used for PMF. Calibration of the instrument was performed externally with $[M + H]^+$ ions of angiotensin I, angiotensin II, substance P, bombesin, and adrenocorticotrophic hormones (clip 1-17 and clip 18-39). Each spectrum was produced by accumulating data from 200 consecutive laser shots. Those samples which were analysed by PMF from MALDI-TOF were additionally analysed using LIFT-TOF/TOF MS/MS analysis. In the TOF1 stage, all ions were accelerated to 8 kV under conditions promoting metastable fragmentation. After selection of jointly migrating parent and fragment ions in a timed ion gate, ions were lifted by 19 kV to high potential energy in the LIFT cell. After further acceleration of the fragment ions in the second ion source, their masses could be simultaneously analysed in the reflector with high sensitivity. PMF and uninterrupted LIFT spectra were interpreted primarily with the Mascot software (Matrix Science Ltd, London, UK). Database searches, through Mascot, using combined PMF and MS/MS datasets were performed via BioTools 2.2 software (Bruker Daltonics). A mass tolerance of 25 ppm and one missing cleavage site for PMF and MS/MS tolerance of 0.5 Da and one missing cleavage site were allowed and oxidation of methionine residues was considered.

Unmatched peaks in the MASCOT database search for HP Q9Y512 were further analysed by de novo sequencing analysis. MS/MS spectra were sequenced de novo using BioTool 2.2 software with full de novo sequencing extension and the top high scoring candidate sequences for MS/MS spectra were then submitted to MS BLAST sequence similarity search (<http://dove.embl-heidelberg.de/Blast2/msblast.html>), which was based on the most likely de novo sequences. Protein identification significance was judged using the MS BLAST scoring algorithm (Shevchenko et al., 2001).

2.3. Results from the chemical analysis of HPs

In Table 1, eight HPs are described providing accession number, protein name, number of identified spots, theoretical molecular weight, theoretical and observed isoelectric point as well as results from data base searches for protein identification.

Fig. 2 illustrates a typical Mascot search result of HP Q9BTR7 indicating the probability based MOWSE score and descriptive details.

Supplementary table lists peptide sequences, matched peptide masses of eight HPs and MS/MS-results of HP Q9Y512. This documentation is essential for reproducibility of identification and is provided as supplementary data.

Fig. 3 demonstrates a typical MS/MS spectrum obtained from peptides resulting from tryptic cleavage unambiguously identifying HP Q9Y512.

In order to increase sequence coverage and confidence of MS identification digestion of HP Q9Y512 was digested with three different proteases.

Trypsin digestion lead to a sequence coverage of 63% that was increased to 67% by digestion with Lys-C and finally to 76% when HP Q9Y512 was cleaved by Asp-N.

In Fig. 4, obtained sequence coverages are graphically demonstrated.

De novo sequencing based upon MS/MS and corresponding de novo sequencing software confirmed identification of HP Q9Y512 by generating the peptide sequence 168–180 (ETSYGISFFQPR).

In Fig. 5, the de novo sequencing results from the LIFT (MS/MS) spectrum of m/z 1431.75 with consensus sequence tags along with scores are shown. In this figure also the MS BLAST search of this peptide is given and highly identical (high score) to Q9Y512.

3. Data mining

3.1. Methods for protein identification from mass spectrometrical data/spectra

The combination of mass spectrometry and protein sequence database searching is an effective method to identify proteins from proteomics experiments (for review: Gras et al., 1999). A key issue in using this method is how to evaluate the quality of the database search results and significance-testing methods (Eriksson et al., 2000; Perkins et al., 1999; Tang et al., 2000; Berndt et al., 1999) have been used to address this task previously. These methods classify results into two categories: significant or non-significant, based on user-set criteria—whether a search is significant depends on criteria selected (subjectively) by the user. To evaluate search results objectively, the method should be largely independent of decision criteria or choice preference (such as stringent, moderate, or liberal). Signal detection theory provides a proper framework for objective evaluation of database search results (Zhang and Chait, 2000).

A number of systems are currently available that claim to allow fast and accurate identification of proteins using mass

Table 1
Mass spectrometrical results of hypothetical proteins in MB cell line (DAOY)

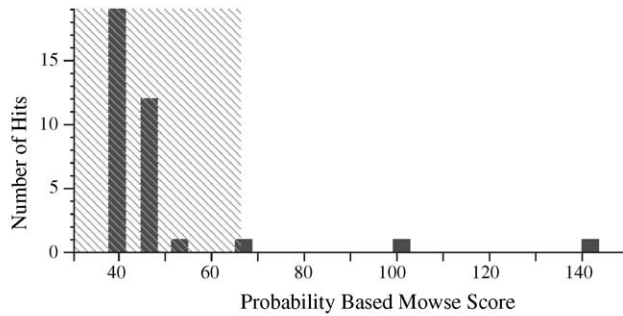
Accession number	Protein name	Number of identified spots	TMW ^a (Da)	TIP ^b	OIP ^c	Mascot-search results (combined MS and MS/MS)					MS-Fit			Profound			
						OMW ^d (Da)	Score	Number of matched peptides	Sequence coverage (%)	Expectation	Identified protein	MOWSE-score	Sequence coverage (%)	Identified protein	Probability	Est' d Z	Seq coverage (%)
Q9BTR7	FLJ20331 protein [Fragment]	1	65980	5.6	6.3	66451	142	26	63	1.5e−009	FLJ20331 protein	4.969e+09	31	FLJ20331 protein (AAH03407.2)	1.0e+000	1.66	40
Q8IY67	RAVER1	1	63877	8.8	6.8	64464	134	24	49	9.2e−009	RAVER1	1.365e+10	29	RAVER1 (AAH37428.1)	1.0e+000	2.33	36
Q96GS8	Hypothetical protein ABCF3	1	79745	5.6	6.0	80094	94	17	37	7.2e−005	Hypothetical protein ABCF3	7.928e+08	36	Similar to hypothetical protein ABCF3 (BAC03881.1) (86% identity)	1.0e+000	2.24	33
Q9Y512	SAM50-like protein CGI-51	1	51962	6.4	6.7	52271	545	33	63	3.3e−050	SAM50-like protein CGI-51	7.728e+14	63	SAM50-like protein CGI-51 (Q9Y512)	1.0e+000	2.36	63
Q96I26	PTPN11 protein	1	5837	9.7	8.5	5890	74	4	48	0.006	Not significant	–	–	–	–	–	–
Q9P0J3	Putative 55 kDa protein	1	55003	7.0	6.8	55481	87	19	42	0.00035	Putative 55 kDa protein	5.978e+08	41	HSCP117 (AAF29081.1)	1.0e+000	2.15	41
Q9JJG2	Mus musculus brain cDNA, clone MNCb-2622, similar to AB033102 KIAA1276 protein	2	68750	5.6	6.0	69392	68	13	20	0.025	Not significant	–	–	Not significant	–	–	–
					6.8		93	18	40	7.9e−005							
Q9BT99	Ras association (RalGDS/AF-6) domain family 5, isoform D	1	43880	9.3	8.0	44536	67	10	30	0.032	Not significant	–	–	Not significant	–	–	–

^a Theoretical molecular weight.

^b Theoretical isoelectric point.

^c Observed isoelectric point.

^d Observed molecular weight.



Q9BTR7 Mass: 66451 Score: 142 Expect: 1.5e-009 Peptides matched: 26

FLJ20331 protein [Fragment]

Sequence Coverage: **63%**
Matched peptides shown in **Bold**

1 DCGTVPQGLLKAARKSGQLNLSGRNLSEVPQCVWRINVDIPEEANQNLS
51 FGATERWWEQTDLTKLIISNNKLQSLTDDLRLLPALTVLDIHDNQLTSLP
101 SAIRELENLQKLNVSHNKLLKILPEEITNLRNLKCLYLQHNELTCISEGFE
151 QLSNLEDLDLNNHLTTVPASFSSLSL VRLNLSSNELKSLPAEINRMKR
201 LKHLDCNSNLEETIPPELAGMESLELLYLRRNKLRFLEPFSPCSLLKELH
251 VGENQIEMLEAEHLKHLNSILVLDLRDNKLSVPDEIILLRSLERLDLSN
301 NDISSLPYSLGNLHLKFLALEGNPLRTIRREIISKGTQEVLYLRSKIKD
351 DGPSQSESATETAMTLPSESRVNIHAITLKLIDYSDKQATLIPDEVFDA
401 VKSNIVTSINFKNQLCEIPKRMVELKEMVSDVDLSFNKLSFISLELCVL
451 QKLTFDLDRNNFNLNLSPEEMESLVRQLTINLSFNRFKMLPEVLYRIFTLE
501 TILISNNQVGSVDPPQKMKMMENLTTLDLQNNDDLQIPPELGNCVNLRTLL
551 LDGNPFRVPRAAILMKGTAAILLEYLRDRIPT

Fig. 2. Presentation of Mascot search result on HP Q9BTR7. Probability Based Mowse Score revealed significant identification by MS.

spectra, such as Mascot, Protein Prospector including MS-Fit, ProFound and SEQUEST, amongst others.

3.1.1. Mascot searches

Mascot is a powerful search engine using mass spectrometry data to identify proteins from primary sequence databases (Perkins et al., 1999). Algorithms (Henzel et al., 1993; Yates et al., 1993) rank proteins according to the decreasing number of matching peptides and an updated method (James et al., 1993) uses a probability-based algorithm. An elegant ranking method proposed by Perkins et al. (1999) combines the frequency distribution of the matching peptides and provides a normalised probability related score, MOWSE (Fig. 2).

Three different Mascot-search methods can be categorised:

- (1) Peptide mass fingerprinting, in which the only experimental data are peptide mass values,
- (2) Sequence query, in which peptide mass data are combined with amino acid sequence and information on amino acid composition and
- (3) MS/MS ion search, using uninterpreted MS/MS data from one or more peptides

The application of the MASCOT database led to the identification of all the eight HPs described herein (Table 1).

3.1.2. Protein Prospector searches

Protein Prospector searches (Clauser et al., 1999) consist of a series of subprogrammes as MS-Fit for peptide-mass fingerprinting, MS-Tag for fragment-ion tag data searches in MS/MS, MS-Seq for sequence tag data searches and MS-Pattern for Edman microsequence data searches. The name *MS-Fit* stems from the program's expected usage: correlating mass spectrometry data (parent masses only, not fragment masses) with a protein in a sequence database which best fits the data. Note that the word fit was chosen and NOT the word identify. In the spring of 1995, when the name was selected, the typical peptide mass fingerprinting experiment preceding use of MS-Fit was to digest a protein with an enzyme, then perform MALDI mass spectrometry on the resulting mixture of peptides to determine the masses of each peptide. At that time the state of the art mass accuracy using MALDI on a continuous-extraction, reflector time-of-flight instrument was ± 0.5 Da. This mass accuracy level was poor in comparison to the standard of ± 10 ppm established with magnetic sector instruments several decades earlier. Thus in our opinion MS-Fit could in favourable cases (where both species and approximate intact protein molecular weight were known) merely *suggest protein identity*. To *establish protein identity* one needed some sequence support. This support could be obtained from the combined use of MS/MS and a

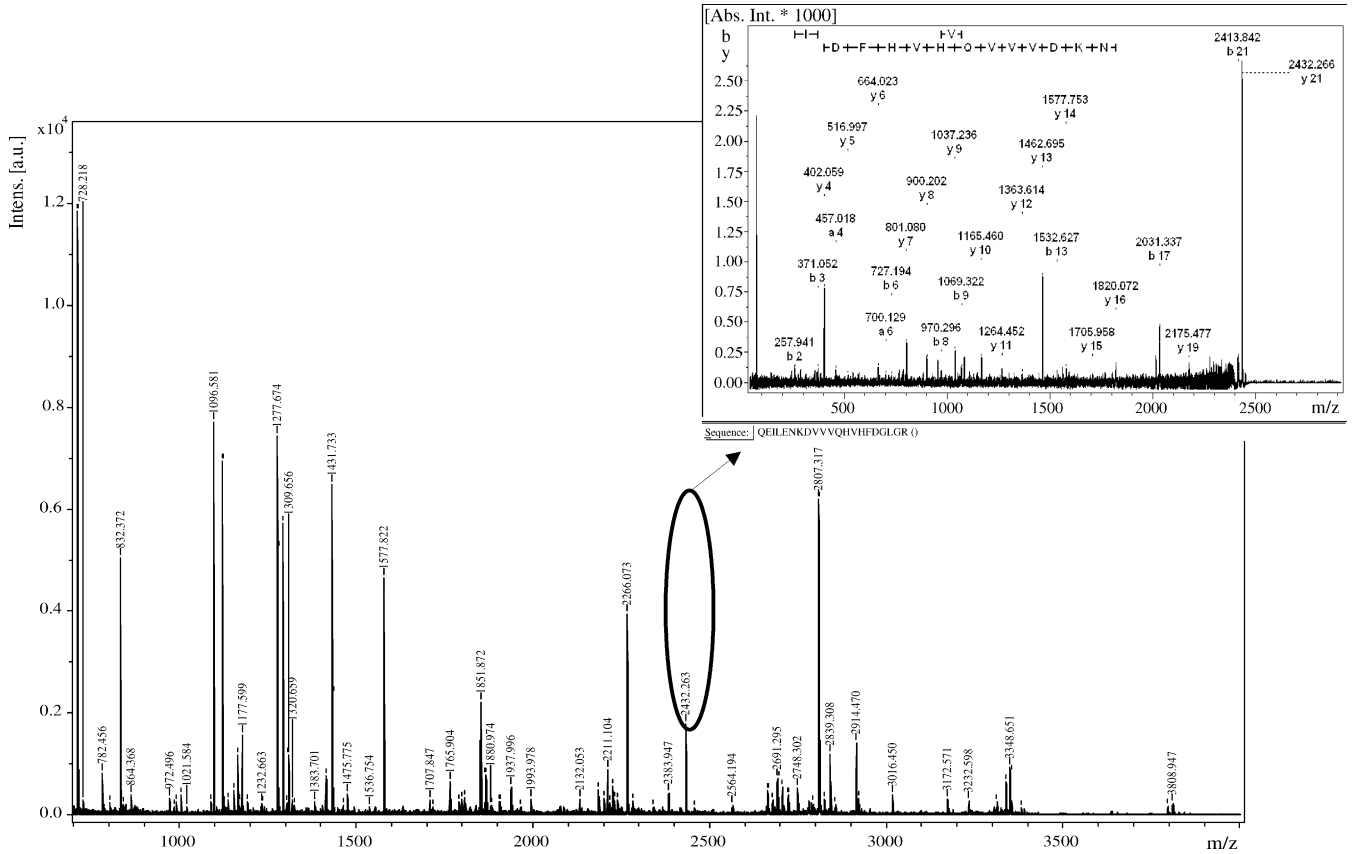


Fig. 3. A typical MS spectrum from tryptic cleavage of Q9Y512 and MS/MS spectrum from peptide QEILENKDVVVQHVFHFDGLGR with the *m/z* of 2432.263 are shown, unambiguously identifying this HP.

subsequently developed program MS-Tag. Consequently, proteins can now be confidently identified by peptide mass fingerprinting using masses alone with MS-Fit. Identification certainty is primarily a function of the level of mass accuracy.

Data mining of five out of eight HPs using MS-Fit and Mascot revealed identical HPs (Table 1). When a short protein sequence with the theoretical mass of 5837 kDa was analysed by MALDI-TOF followed by MS-Fit, no significant result was obtained, most probably due to the short length although

```

MGTVHARSLE PLPSSGPDFG GLGEEAEFVE VEPEAKQEIL ENKDVVVQHV HFDGLGRTK
-----
DDIICEIGDV FKANLIEVM RKSHEAREKL LRLGIFRQVD VLIDTCQGDD ALPNGLDVTF
-----
EVTELRRLTG SYNTMVGNN E GSMVLGLKLP NLLGRAEKVT FQFSYGTKET SYGLSFFKPR
-----
PGNFERNFSV NLYKVTGQFP WSSLRETRDG MSAEYSFPIW KTSHTVKWEG VWRELGLSR
-----
TASFAVRKES GHSLKSSLSH AMVIDSRNSS ILPRRGALLK VNQELAGYTG GDVSIKEDF
-----
ELQLNKQLIF DSVFSASFWG GMLVPIGDKP SSIADRFYLG GPTSVRGFMS HSIGPQSEGD
-----
YLGGEAYWAG GLHLYTPLPF RPGQGGFGE L FR THFFLNAG NLCNLNYGEG PKAHIRKLAE
-----
CIRWSYGAGI VLRLGNIARL ELNYCVPMGV QTGDRICDV QFGAGIRFL
-----
    
```

Fig. 4. The protein sequence of HP Q9Y512 and individual sequence coverages obtained by proteolytic degradation by trypsin (solid grey line), Lys-C (dot line) and N-Asp (solid black line) are shown that together finally lead to 76% of sequence coverage.

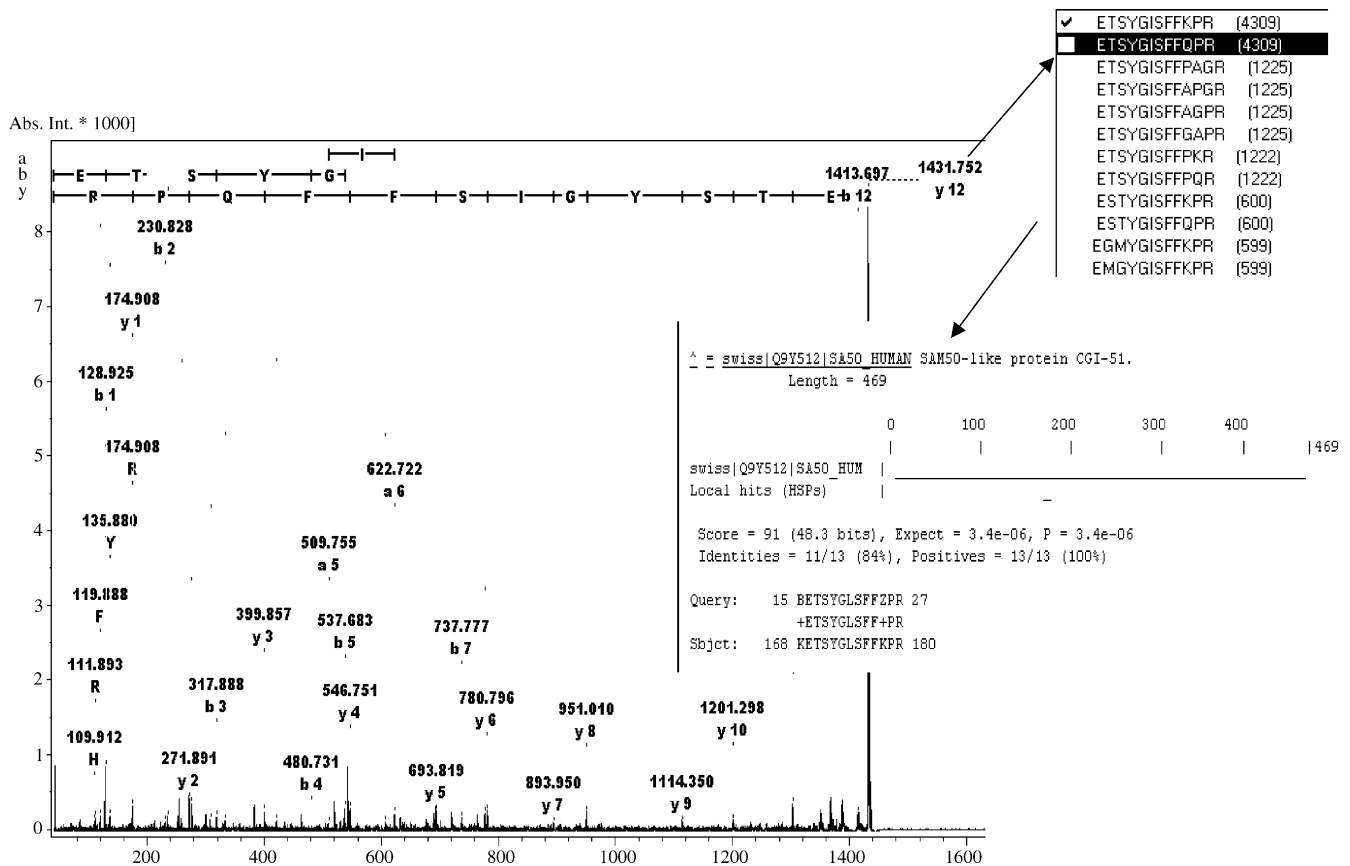


Fig. 5. MS/MS spectrum in the LIFT mode (Suckau et al., 2003) forming the basis for de novo sequencing of HP Q9Y512. The parental peptide with a m/z of 1431.75 was used and the sequence 169–180 (ETSYGLSFFKPR) was generated. The sequence given in letters representing amino acids is incicated on top of the figure. Using the de novo sequencing software and consensus sequence tags the correct sequence is provided and presenting with high score of 4309. The MS BLAST search provided strong evidence for the correct identification of the peptide and assignment to HP Q9Y512.

sequence coverage was 48%. Two out of eight proteins were differently identified by Mascot and MS-Fit: not only experimental data would clarify which database identified the correct protein due to the short sequence.

3.1.3. ProFound searches

ProFound is an expert system for identifying proteins using MS peptide mapping data (Zhang and Chait, 2000). ProFound ranks protein candidates using a Bayesian algorithm that takes into account individual properties of each protein in the database as well as other information relevant to the experiment. The probability for the sample protein to be a specific protein in the database is calculated using the MS data as well as other background information such as the mass range in which the protein is expected to lie, identity of species from which the protein originated, mass accuracy, enzyme cleavage chemistry, protein sequence, previous experiments on the sample protein, to name a few.

Based upon MS data ProFound identified five out of eight HPs that were identified from Mascot searches (Table 1); these sequence were significantly annotated when the Z-score was respected (Eriksson et al., 2000). These five identifications were in agreement with Mascot search results and in the case of Q9POJ3 a protein sequence with 96% identity to HSCP117 was aligned.

Chamrad and co-workers have carried out a comparative study on the use of several protein identification algorithms for the identification of known proteins (Chamrad et al., 2004); it would now be important to systematically perform comparative studies for the application of these databases using HPs.

Although one may expect that combination of the three abovementioned data bases may lead to identification of more proteins than by one database only, the current data on HPs show that this is not the case; even contradictory results may be obtained, an observation reported by Marcus et al. (2000). This subject is still open and has to be addressed in further studies.

3.2. Alignments and BLAST searches

3.2.1. Alignments

The comparison of nucleotide or protein sequences from the same or different organisms is a powerful tool in protein chemistry. By finding similarities between sequences, scientists can infer the function of newly sequenced genes, predict new members of gene families, and explore evolutionary relationships.

There are two main types of alignments: global and local. A global alignment such as given in the method of Needleman and Wunsch (1970) contains the entire sequence of each protein.

One of the first important algorithms for globally aligning two protein sequences was described by Needleman and Wunsch (1970). This algorithm is important because it produces an optimal alignment of proteins, even allowing the introduction of gaps. The Needleman–Wunsch alignment approach can be described in three steps: (1) setting up a matrix, (2) scoring the matrix, (3) identifying the optimal alignment (Needleman and Wunsch, 1970). This is an example for dynamic programming algorithm. It is called “dynamic” because the alignment is created on a residue-by-residue basis in a search for optimal alignment. The programming refers to the use of a set of rules to determine the alignment.

As to local alignments, the local alignment algorithm described by Smith and Waterman (1981) is the most rigorous method by which subsets of two protein sequences can be aligned and FASTA includes both, global and local algorithms.

FASTA (<http://www.ebi.ac.uk/fasta/>) is a currently widely-used database.

This new version of FASTP (Lipman and Pearson, 1985), was described by Pearson and Lipman (1988). It uses an improved algorithm that increased sensitivity (detection of distantly related sequences) with a small loss of selectivity (similarity score of unrelated sequences) and a negligible decrease in speed. FASTA stands for FAST-All, reflecting the fact that it can be used for fast protein comparison. This program achieves a high level of sensitivity for similarity searching at high speed that is achieved by performing optimised searches for local alignments using a substitution matrix. The high speed of this program is achieved by using the observed pattern of word hits to identify potential matches before attempting the more time consuming optimised search. FASTA can be very specific when identifying long regions of low similarity especially for highly divergent sequences. Sequence similarity searching against complete proteomes or genome databases can be performed.

FASTA includes an additional step in the calculation of the initial pairwise similarity score that allows multiple regions of similarity to be joined to increase the score of related sequences. The original LFASTA program was able to display all the regions of local similarity between two sequences with scores greater than threshold, using the same scoring parameters and a similar alignment algorithm and these local similarities can be displayed as a “graphic matrix” plot or as individual alignments (Pearson and Lipman, 1988).

For basic similarity searches FASTA3 is a useful tool (<http://www.ebi.ac.uk/fasta33/>).

In the case of our eight HPs all structures were aligned with FASTA3 and identities (% amino acid residues identical) to known proteins were between 25 and 31% (Table 2). A typical FASTA3 alignment (combined local and global) is provided for HP Q9BTR7 (Fig. 6).

Similarities (percentage of similar and identical matches) were calculated in the range between 45 and 71%. This relatively high similarities do not contradict the term HP as similarities calculated were similar to proteins that were HPs as well (Table 2).

Results in Table 2 are expressed as scores at various stages of alignment, *initn*: the highest scoring region at the end of the first alignment; *initl*: the highest score at the end of the last alignment; *opt*: optimal alignment score; *Z-score*: Z is used to indicate the likelihood that a candidate belongs to a random match population in the sense of traditional statistics; *bits*: the score of a local alignment; *E*(): is the expectation value lower limit allowing to focus on more distant relationships. *Smith–Waterman score*: this score represents the outcome of a local alignment respecting errors by comparing high sequences.

3.2.2. BLAST searches

The Basic Local Alignment Search Tool (BLAST) (Altschul et al., 1990, 1997) is the tool most frequently used for calculating sequence similarity. BLAST comes in variations for use with different query sequences against different databases. All BLAST applications, as well as information on which BLAST program to use and other help documentation, are listed on the BLAST homepage (<http://www.ncbi.nlm.nih.gov/BLAST/>).

There are many different variations of BLAST available for different sequence comparisons, e.g. a DNA query to a DNA database, a protein query to a protein database (BLAST2, Blastp search), and a DNA query, translated in all six reading frames, to a protein sequence database. Other adaptations of BLAST, such as PHI-BLAST (pattern hit initiated blast), PSI-BLAST (for iterative protein sequence similarity searches using a position-specific score matrix) and RPS-BLAST (for searching for protein domains in the conserved Domains Database) perform comparisons against sequence profiles.

BLAST performs “local” alignments. Most proteins are modular in nature, with functional domains often being repeated within the same protein as well as across different proteins from different species. The BLAST algorithm is tuned to find these domains or shorter stretches of sequence similarity.

Standard protein–protein BLAST (blastp) is used for both, identifying a query amino acid sequence and for finding similar sequences in protein databases. Like other BLAST programs, blastp is designed to find local regions of similarity. When sequence similarity spans the whole sequence, blastp will also report a global alignment, which is the preferred result for protein identification purposes (Altschul et al., 1997; <http://www.ncbi.nih.gov/BLAST>).

Blastp searches for our HPs revealed high similarity to hypothetical proteins of unknown function (Table 2). For putative 55 kDa protein e.g. the Blastp result is confirming 96% identity and calculated 96% Positives (conservative substitutions). No Gaps were identified.

BLAST2 takes two input sequences and compares them directly producing gapped alignments. “Aligning two sequences” regards the longer, second sequence as the database. Unlike the other BLAST programs, there is no need to format the database sequence in any special way. Since translated BLAST programs are incorporated in this program, the second sequence can be of different type as long as an appropriate BLAST program is selected (Bimpikis et al., 2003).

Table 2
Protein sequence alignment of hypothetical proteins in MB cell line (DAOY)

Accession number	Protein name	BLAST2	PHI-BLAST	PSI-Blast	RPS-Blast	Blastp search (standard)	MP search	FASTA
Q9BTR7	FLJ20331 protein [Fragment]	Leucine-rich repeat protein; contains similarity to elicitor-inducible receptor EIR (Q9LUQ2)	Similar to Rho-GTPase-activating protein 7 (Deleted in liver cancer 1 protein) (Dlc-1) (Q96QB1)	Leucine-rich repeat protein SHOC-2 (Ras-binding protein Sur-8) (O88520)	gn1 CDD 5333; PSSM-Id: 5333; cd00116, LRR_RI, Leucine-rich repeats (LRRs), ribonuclease inhibitor (RI)-like subfamily	Leucine-rich repeat protein SHOC-2 (Ras-binding protein Sur-8) (O88520)	Leucine-rich repeat protein SHOC-2 (Ras-binding protein Sur-8) (O88520)	Leucine-rich repeat protein; contains similarity to elicitor-inducible receptor EIR (Q9LUQ2)
		Species: <i>Arabidopsis thaliana</i>	Species: mouse	Species: mouse	Species: mouse	Species: mouse	Species: mouse	Species: <i>Arabidopsis thaliana</i>
		253 bits (709)	Length: 1091	Score: 140 bits (353)	Score: 39.5 bits (92)	142 bits (358)	Score: 486	Sequence length: 594 aa
		<i>E</i> -value: 4e−66	Score: 19.2 bits (38)	Expect: 8e−33	CD-Length: 319 residues	<i>E</i> -value: 2e−33	Match 35.1%	initn: 823
		Identities: 210/616 (34%)	Expect: 8.5	Identities: 138/452 (30%)	77.4% aligned	Identities: 128/497 (25%)	QryMatch 9.7%	initl: 305
		Positives: 329/616 (53%)	Identities: 8/34 (23%)	Positives: 221/452 (48%)	Expect: 0.001	Positives: 216/497 (43%)	Pred. No. 2.99e−86	opt: 796
		Gaps: 72/616 (11%)	Positives: 15/34 (44%)	Gaps: 26/452 (5%)		Gaps: 78/497 (15%)	Matches 97	Z-score: 881.8
							Conservative 65	bits: 173.1
							Mismatches 106	<i>E</i> (): 4.8e−41
							Indels 8	35.113%
							Gaps 7	identity (40.037% ungapped) in 618 aa overlap (10-580:4-592)
Q8IY67	RAVER1	Proline/glutamine rich splicing factor (Q6DRK8)	Similar to Caspase recruitment domain protein 10 (Bcl10-interacting MAGUK protein 1) (Bimp1) (P58660)	Similar to Splicing factor, proline-and glutamine-rich (P23246)	gn1 CDD 25314; smart00360, RRM, RNA recognition motif	Splicing factor, proline-and glutamine-rich (P23246)	Splicing factor proline/glutamine rich (Q86VG2)	Alpha 3 collagen IV (Q9QZS0)
		Species: <i>Brachydanio rerio</i>	Species: human	Species: mouse	Species: mouse	Species: human	Species: mouse	Species: mouse
		Score: 68.5 bits (180)	Score: 21.9 bits (45)	Score: 65.9 bits (159)	CD-Length: 73 residues	Score: 71.2 bits (173)	Score 246	Sequence length: 1699
		Expect: 3e−10	Expect: 1.3	Expect: 3e−09	95.9% aligned	Expect: 7e−12	Match 32.6%	initn: 119
		Identities: 50/184 (27%)	Identities: 15/48 (31%)	Identities: 44/139 (31%)	Score: 65.3 bits (159)	Identities: 44/139 (31%)	QryMatch 4.7%	initl: 74
		Positives: 86/184 (46%)	Positives: 20/48 (41%)	Positives: 70/139 (50%)	Expect: 2e−11	Positives: 70/139 (50%)	Pred. No. 2.46e−28	opt: 262
		Gaps: 7/184 (3%)		Gaps: 2/139 (1%)		Gaps: 2/139 (1%)	Matches 45	Z-score: 227.7 bits: 53.7
							Conservative 32	<i>E</i> (): 0.00013
							Mismatches 58	Smith-Waterman score: 262

							Indels 3 Gaps 3	30.556% identity (34.241% similar) in 288 aa overlap; (256-529:911-1181)
Q96GS8	Hypothetical protein ABCF3	ATP-binding cassette, sub-family F (GCN20), member 3 (Q8K268) Score: 1236 bits (3512) Expect: 0.0 Identities: 687/709 (96%) Positives: 701/709 (97%)	Lethal factor precursor (LF) (Anthrax lethal toxin endopeptidase component) (P15917) Score: 25.0 bits (53) Expect: 0.22 Identities: 16/66 (24%) Positives: 31/66 (46%) Gaps: 4/66 (6%)	Isoform of ABCF3 protein (fragment) [ABCF3] (Q86UA2) Score: 1395 bits (3611) Expect: 0.0 Identities: 709/709 (100%) Positives: 709/709 (100%)	1gn CDD 5341; cd00267, ABC_ATPase, ABC (ATP-binding cassette) transporter nucleotide-binding domain CD-Length: 217 residues 97.2% aligned Score: 115 bits (290) Expect: 2e-26	ATP-binding cassette, sub-family F, member 1 (Q6P542) Species: mouse Score: 372 bits (954) Expect: e-102 Identities: 215/570 (37%) Positives: 309/570 (54%) Gaps: 27/570 (4%)	ATP-binding cassette, sub-family F (GCN20), member 3 (Q8K268) Species: mouse Score 5974 Match 96.9% QryMatch 97.4% Pred. No. 0.00e+00 Matches 687 Conservative 16 Mismatches 6 Indels 0 Gaps 0	Abcf3 protein (Q8K268) Species: mouse (709 aa) initn: 4495; init1: 4495 opt: 4495 Z-score: 4194.7 bits: 786.7 E(): 0 Smith-Waterman score: 4495 96.897% identity (96.897% ungapped) in 709 aa overlap (1-709:1-709)
Q9Y512	SAM50-like protein CGI-51	Mitochondrion protein, putative (Q5KEW4) Score: 134 bits (368) Expect: 3e-30 Identities: 115/382 (30%) Positives: 194/382 (50%) Gaps: 22/382 (5%)	Similar to Caspase recruitment domain protein 10 (Bcl10-interacting MAGUK protein 1) (P58660) Score: 21.2 bits (43) Expect: 2.0 Identities: 14/39 (35%) Positives: 18/39 (46%) Gaps: 3/39 (7%)	Similar to Outer membrane protein (Q6G1J3) (Bartenolla quintana) Score: 87.4 bits (215) Expect: 7e-16 Identities: 127/488 (26%) Positives: 206/488 (42%) Gaps: 68/488 (13%)	gn1 CDD 25688; pfam01103, Bac_surface_Ag, Surface antigen CD-Length: 313 residues 99.7% aligned Score: 187 bits (477) Expect: 2e-48	Sorting assembly machinery protein 50 (P53969) Species: Yeast Score: 103 bits (258) Expect: 7e-22 Identities: 73/226 (32%) Positives: 116/226 (51%) Gaps: 15/226 (6%)	Similar to Mitochondrion protein, putative (Q5KEW4) Score 293 Match 31.4% QryMatch 7.1% Pred. No. 6.38e-49 Matches 116 Conservative 82 Mismatches 152 Indels 20 Gaps 17	Putative outer membrane transmembrane protein (Q92Q48) initn: 153 init1: 86 opt: 367 Z-score: 437.2 bits: 90.9 E(): 2.8e-16 Smith-Waterman score: 367 25.408% identity (26.716% ungapped) in 429 aa overlap (47-468:362-776)
Q96I26	PTPN11 protein	- ? Blast for short sequences: hypothetical protein F47D12.3 in chromosome III (Q09389)	Similar to hypothetical 73.7 KD protein Y4FB (P55440)	Similar to growth regulator-related protein [T18N24.40] (<i>Arabidopsis thaliana</i> (Mouse-ear cress))	-	Limbin (Q86UK5)	Hypothetical protein ECU02 (Q8SWA7)	Hypothetical protein ECU02 (Q8SWA7)

Table 2 (Continued)

Accession number	Protein name	BLAST2	PHI-BLAST	PSI-Blast	RPS-Blast	Blastp search (standard)	MP search	FASTA
		Score: 34.6 bits (74) Expect: 0.058 Identities: 22/39 (56%) Positives: 24/39 (61%) Gaps: 11/39 (28%)	Score: 19.6 bits (39) Expect: 0.35 Identities: 7/17 (41%) Positives: 12/17 (70%)	Score: 31.2 bits (69) Expect: 6.5 Identities: 14/42 (33%) Positives: 24/42 (57%) Gaps: 1/42 (2%)		Score: 30.0 bits (66) Expect: 1.4 Identities: 13/38 (34%) Positives: 23/38 (60%)	Score 113 Match 32.6% QryMatch 24.5% Pred. No. 8.55e-07 Matches 14 Conservative 13 Mismatches 15 Indels 1 Gaps 1	initn: 86 init1: 61 opt: 96 Z-score: 153.8 bits: 33.2 E(): 1.7 Smith-Waterman score: 96 33.333% identity (34.146% ungapped) in 42 aa overlap (7-47:126-167)
Q9P0J3	Putative 55 kDa protein	P55 (Q6AYT3) Species: rat Score: 987 bits (2552) Expect: 0.0 Identities: 491/505 (97%) Positives: 491/505 (97%)	Similar to Glucan endo-1,3-beta-glucosidase, acidic isoform precursor ((1->3)-beta-glucan endohydrolase) (P49237) Score: 23.5 bits Expect: 0.37 Identities: 13/37 (35%) Positives: 17/37 (45%) Gaps: 9/37 (24%)	P55 (Q6AYT3) Species: rat Score: 1016 bits (2627) Expect: 0.0 Identities: 494/505 (97%) Positives: 496/505 (98%)	gn1 CDD 16900 pfam01139, UPF0027, Uncharacterized protein family UPF0027 CD-Length: 436 residues 100.0% aligned Score: 551 bits (1420) Expect: 8e-158	P55 (Q6AYT3) Species: rat Score: 980 bits (2534) Expect: 0.0 Identities: 487/505 (96%) Positives: 489/505 (96%)	P55 (Q6AYT3) Species: rat Sequence length: 505 Match percent: 98 Match query percent: 97.7 Score: 4295 E-value: 0.00e+00	P55 (Q6PX76) Species: rat initn: 3282 init1: 3282 opt: 3282 Z-score: 3755.6 bits: 704.4 E(): 4.1e-201 Smith-Waterman score: 3282 97.822% identity (97.822% ungapped) in 505 aa overlap (1-505:1-505)
Q9JG2	Mus musculus brain cDNA, clone MNCb-2622; Similar to AB033102 KIAA1276 protein	Nuclear mitotic apparatus protein 1 (Q14980) (human) Score: 88.6 bits (218) Expect: 3e-16 Identities: 96/479 (20%) Positives: 201/479 (41%) Gaps: 23/479 (4%)	Similar to Calmodulin-sensitive adenylate cyclase precursor (P40136) Score: 23.5 bits (49) Expect: 0.37 Identities: 13/37 (35%) Positives: 17/37 (45%) Gaps: 9/37 (24%)	Similar to myosin heavy chain, skeletal muscle, adult 2 (Q9UKX2) (human) Score: 52.0 bits (123) Expect: 4e-05 Identities: 66/248 (26%) Positives: 120/248 (48%) Gaps: 10/248 (4%)	gn1 CDD 16998; pfam01576, Myosin_tail_1, Myosin tail CD-Length: 860 residues Only 28.7% aligned Score: 41.5 bits (97) Expect: 3e-04	Nuclear mitotic apparatus protein 1 (Q14980) (human) Score: 88.6 bits (218) Expect: 4e-17 Identities: 96/479 (20%) Positives: 201/479 (41%) Gaps: 23/479 (4%)	Medulloblastoma antigen MU-MB-20.201 [Fragment] (Q7Z5E6) Sequence length: 681 Match percent: 34.9 Query percent: 3.5 Score: 181 E-value: 2.00e-11	Myosin heavy chain, smooth muscle isoform (P35749) species:human initn: 190 init1: 89 opt: 307 Z-score: 266.3 bits: 60.1 E(): 9.2e-07

Q9BT99	Ras association (RalGDS/AF-6) domain family 5, isoform D	Ras effector-like protein (Q8WXF4)	Similar to lethal factor precursor (Anthrax lethal toxin endopeptidase component)	Similar to Ras effector-like protein (Q8VXF4)	gn1 14827; cd00153, RA, RasGTP binding domain from guanine nucleotide exchange factors	Protein kinase C, D2 type (Q9BZL6)	Ras effector-like protein (Q8WXF4)	Smith-Waterman score: 307 23.841% identity (27.586% ungapped) in 604 aa overlap (35-577:194-776)
		Score: 344 bits (968)	Score: 28.9 bits (63)	Score: 362 bits (929)	CD-Length: 89 residues 100.0% aligned Score: 65.0 bits (158)	Score: 42.4 bits (98)	Score 2796	initn: 1254
		Expect: 1e-93 Identities: 189/191 (98%) Positives: 191/191 (99%)	Expect: 0.007 Identities: 20/75 (26%) Positives: 41/75 (54%) Gaps: 3/75 (4%)	Expect: 8e-99 Identities: 181/191 (94%) Positives: 187/191 (97%)	Expect: 2e-11	Expect: 0.002 Identities: 19/56 (33%) Positives: 29/56 (51%) Gaps: 3/56 (5%)	Match 84.7% QryMatch 80.6%	init1: 1220 opt: 1228
							Pred. No. 0.00e+00 Matches 322 Conservative 25 Mismatches 30 Indels 3 Gaps 3	Z-score: 1141.2 bits: 219.4 E(): 1.7e-55 Smith-Waterman score: 1228 98.438% identity (98.438% ungapped) in 192 aa overlap (189-380:38-229)

Leucine-rich repeat protein (Q9LUQ2)
 Sequence length: 594 aa
 initn: 823 init1: 305 opt: 796 Z-score: 881.8 bits: 173.1 E(): 4.8e-41
 Smith-Waterman score: 928; 35.113% identity (40.037% ungapped) in 618 aa overlap (10-580:4-592)

```

10 20 30 40 50
Sequen DCGTSVPQGLLKAARKSGQLNLSGRNLSEVPQCVWRINVDIPEEANQNL-SFGATERWWE
      :::::::::::  :::::::::::  :::::::::::  :::::::::::  :::::::::::
UNIPRO  MDRILKAARTSGSLNLSNRSLK-----DVPTEVYQCLETTGEGENWWE
      10 20          30 40

60 70 80 90 100 110
Sequen QTDLTKLIISNNKLSLTDLLRLLPALTVLDIHDNQLTSLPSAIRELENLQKLNVSHNKL
      :::::::::::  :::::::::::  :::::::::::  :::::::::::  :::::::::::
UNIPRO  AVDLQKLILAHNDIEVLRDLKLNLAQLVVLNVSHNKLSQLPAAIGELTAMKSLDVSFNLS
      50 60 70 80 90 100

120 130 140 150 160 170
Sequen KILPEEITNLRNLKCLYLQHNETLCISEGFQELSNLEDLDLSNNHLTVPASFSLSLW
      :::::::::::  :::::::::::  :::::::::::  :::::::::::  :::::::::::
UNIPRO  SELPEQIGSAISLVKLDLSSNRLKELPDSIGRCLDLSDLKATNNQISSLPEDMVNCSKLS
      110 120 130 140 150 160

180 190 200 210 220 230
Sequen RLNLSSNELKSLPAE-INRMKRLKHLDCNSNLETPPELAGMESLELLYLRRNKLRLFP
      :::::::::::  :::::::::::  :::::::::::  :::::::::::  :::::::::::
UNIPRO  KLDVEGNKLTALSENHIASWTMLAELNACKNMLGVLPQNIQSLRRLRLDLHQNKISSVP
      170 180 190 200 210 220

240 250 260 270 280 290
Sequen EFPS---CSLLKELHVGGENQIEMLEAEHLKHLNSILVLDLRDNKLSVPDEIILLRSLER
      :::::::::::  :::::::::::  :::::::::::  :::::::::::  :::::::::::
UNIPRO  --PSIGGCSSLVEFYLGINSLSLTPAE-IGDLSRLGTLDLRSNQLKEYPVGACKLK-LSY
      230 240 250 260 270

300 310 320 330 340 350
Sequen LDLSNNDISSLPYSLGNLH-LKFLALEGNPLRTIRREIISKGTQEVLYLRSKIKDDGSPS
      :::::::::::  :::::::::::  :::::::::::  :::::::::::  :::::::::::
UNIPRO  LDLSNNSLTGLHPELGNMTTLRKLVLVGNPLRTLRSLLVNGPTAALLKYLRSLNSNEET
      280 290 300 310 320 330

360 370 380 390 400 410
Sequen QSESATETAMTLPSESRVNIHAI-ITLKILDYSDKQATLIPDEVFDAVKSNIIVTSINFSK
      :::::::::::  :::::::::::  :::::::::::  :::::::::::  :::::::::::
UNIPRO  SASTPTKENV-IASAARMSISSKELSLEGLNLSD----VPSEVWE---SGEITKVNLSK
      340 350 360 370          380 390

420 430 440 450 460
Sequen NQLCEIPKRM---VELKEMVSDVDLSFNKLS-FISLELCVLQKLTFLDLRNNFLNSLPEE
      :::::::::::  :::::::::::  :::::::::::  :::::::::::  :::::::::::
UNIPRO  NSIEELPAQLSTSVSLQTLI----LSRNKIKDWPGAILKSLPNLMCLKLDNNPLNQIPLD
      400 410 420 430 440

470 480 490 500
Sequen MESLVR-LQTINLSFN-----RFKMLPEVL-----YRI-FTL-----ETIL
      :::::::::::  :::::::::::  :::::::::::  :::::::::::
UNIPRO  GFQVVSGLQILDLSVNAVSVFREHPKFCFLPQLREL YLRYRIPYTSLDSRIQLSEVPEDIL
      450 460 470 480 490 500

510 520 530 540 550
Sequen -ISN-----NQVGSVDPQKMKMMENLTTLDLQNNDDLQIPPELGNCV-NLRTLLLDG
      :::::::::::  :::::::::::  :::::::::::  :::::::::::
UNIPRO  NLSNLIILDLNQNSLQSI-PKGIKNMTSLKHLDISNNNISSLPPPELGLLEPTLEVLRLDG
      510 520 530 540 550 560

560 570 580
Sequen NPFRVPRAAILMKGTAAILYLRDRIPT
      :::::::::::
UNIPRO  NPLRSIRRPILERGTKAVLNYLKDRLPDQ
      570 580 590

```

Fig. 6. A multiple local and global alignment of Q9BTR7 in FASTA3 revealed low identity and similarity to leucine-rich repeat protein Q9LUQ2. Protein name, sequence length, E-value and scores are shown.

Searching our HPs in BLAST2 showed similarity to the same proteins as from blastp, and in the case of putative 55 kDa protein revealed the same protein, with comparable (97%) identities, positives and of course, different scoring (Table 2).

PSI-BLAST is designed for more sensitive protein–protein similarity searches (Altschul et al., 1997). Position-specific iterative (PSI)-BLAST is the most sensitive BLAST program, making it useful for finding very distantly related proteins. PSI-BLAST searches are often used when standard protein–protein BLAST searches either failed to find significant hits, or returned hits with descriptions such as “hypothetical proteins” or “similar to . . .”. The first round of PSI-BLAST is a standard protein–protein BLAST search. The program builds a position-specific scoring matrix which in turn produces a profile designed to identify the key positions of conserved amino acids within a motif. It is able to detect subtle relationships between proteins that are distant structural or functional in nature and that may not have been detected by other BLAST searches.

Searching similarities for our HPs, it was striking that PSI results were contradicting BLAST2 and BlastP results identifying myosin heavy chain rather than nuclear mitotic apparatus protein 1 in BLAST2 and BlastP (Table 2). Higher percentages for identity and positives were observed using PSI-BLAST although different expectations do not allow a final conclusion.

Based upon the fact that profile-based methods as PSI-BLAST clearly outperform other kinds of methods in the low sequence identity realm (Sauder et al., 2000; Altschul and Koonin, 1998), we would trust data from the PSI-BLAST search, however. And indeed, PSI-BLAST searches are even used for novel protein domain families and hypothetical proteins of unknown functions (Doerks et al., 2002; Stevens, 2005).

PHI-BLAST can be used for a restricted protein pattern search (Zhang et al., 1998).

Pattern-hit initiated (PHI)-BLAST is designed to search for proteins that contain a pattern specified by the user and are similar to the query sequence in the vicinity of the pattern. The program takes as input a protein sequence and a pattern of interest in the protein. PHI-BLAST was created to combine pattern search with the search for statistically significant sequence similarity and is a single-pass search method and does not replace, e.g. PSI-BLAST searches.

Unlike other BLAST searches, PHI-BLAST identified hypothetical protein Q9BTR7 as similar to Rho-GTPase-activating protein 7 (Table 2). At the first glance this looks contradictory; searching the presence of leucine-rich repeats (LRR), however, clearly shows that the similar protein contains 11 LRR (Hillig et al., 1999) and therefore the PHI-BLAST search extends similarity searches. *RPS-BLAST*. Reversed position specific BLAST (RPS-BLAST) is a more sensitive way of identifying conserved domains in proteins than standard BLAST searching and represents also a database for sequence superfamilies of protein domains (Pandit et al., 2004). It compares a protein sequence against a database of position specific scoring matrix (PSSM). The PSSMs used in CDD

search capture the substitution frequencies at each position in the multiple sequence alignments of recognized conserved domains. The conserved domain alignments are from the NCBI's CDD, which contains alignments from protein domain databases: SMART, Pfam, COG, and cd (Marchler-Bauer and Bryant, 2004). Search results are displayed as domain architecture cartoons and pairwise alignments between the query and the domain-model consensus sequences.

As for HP Q9BTR7, leucine-rich repeats were detected and an assignment to the ribonuclease inhibitor-like superfamily was predicted (Table 2). Two codes are provided, the gn1 CDD (conserved domain database) 5333 and PSSM—identity: 5333 as well as the CD-length and percentage of residues aligned. The PSI-BLAST results were in agreement with other BLAST and domain searches (below).

As for HP Q96I26, no hit was calculated (Table 2), most probably because of the short sequence (50 amino acids). BLAST search for short sequences (“Search for short, nearly exact matches”) proposed an identity of 56% with 61% positives to hypothetical protein F47D12.3 (Q09389; Table 2). MPsrch is a biological sequence comparison tool that implements the true Smith and Waterman algorithm (see above). It runs a search on a HP/COMPAQ cluster (<http://www.ebi.ac.uk/MPsrch/>), using single and parallelised versions of the software. It allows rigorous searches in a reasonable computational time. MPsrch utilises an exhaustive algorithm, which is recognised as the most sensitive sequence comparison method available, whereas BLAST and FASTA utilise a heuristic one. As a consequence, MPsrch is capable of identifying hits in cases where BLAST and FASTA fail and also reports fewer false-positive hits. MPsrch_pp applies a protein query to search a protein sequence database, using amino acid match scoring derived from a specified table. Only a single gap penalty is used during the search; in most cases, the best alignments between related sequences does not involve long gaps or regions with multiple gaps. Searching for our HPs, we show an example for HP Q8IY67 (RAVER1): MPsrch_pp indicated high and significant similarity to human splicing factor, proline/glutamine-rich (Q86VG2; Table 2). Other databases predicted similarities to different other proline- and glutamine-rich proteins, including collagen IV (alpha3) and RRM, RNA recognition motif. Based on several research outcome it would be probably more reliable to simply proposing “proline- and glutamine-rich proteins” to go further into this direction although all collagens are proline-rich proteins and the core of the message is true but pretends to be more specific. Most search results were compatible with other database predictions.

4. Functional analysis of hypothetical proteins

4.1. Domain analysis–databases

Domains are evolutionary units and most proteins consist of one or more domains. The definition of protein domains varies widely across the discipline of biology although domains are defined simultaneously as:

- (a) regions that display a significant level of sequence homology,
- (b) a minimal part of a gene capable of performing a function,
- (c) a region of the protein with an experimentally assigned function,
- (d) parts of structures that have significant similarity or are compact spatially distinct units of protein structure (Veretnik et al., 2004).

A structural domain may be defined as one or several segments of a polypeptide that forms a compact and stable structure with an associated hydrophobic core and that can fold and function independently of other parts of the sequence. Structural domains are often associated with identifiable cellular/biochemical functions (Orengo et al., 1999).

There may be a limited repertoire (Chothia, 1992; Wolf et al., 2000) of a few thousand domains and only one fifth to one third of known proteins are single-domain structures, the majority are multi-domain proteins (Apic et al., 2003).

There is a large series of domain databases but we have been selecting a couple of widely used programs as PROSITE, PRINTS, InterPro, ProDom, Pfam and SMART.

PROSITE was originally reported by Bairoch and Bucher (1994) and is a method identifying functions of uncharacterised proteins translated from genomic or cDNA sequences. It consists of biologically significant patterns and profiles indicating protein families or known domains (Falquet et al., 2002).

PROSITE failed to provide domain informations for HP Q9Y512, Q96I26 and Q9JG2 showing the limitation of the database for hypothetical structures and only motifs were recognised (Table 3).

PRINTS represents a collection of protein fingerprints that can be used to generate family and tentative functional assignments for uncharacterised sequences (Attwood et al., 2000; Attwood, 2002). This database was of no use for the domain search of our HPs as no domain was indicated for any of them and only a typical leucine rich repeat subtype was observed for Q9BTR7, that was also revealed by other databases (Table 3).

InterPro is an integrated documentation resource of protein families, domains and functional sites and was generated to incorporate major protein signature databases including *PROSITE*, Pfam, PRINTS, ProDom, SMART, TIGRFAMs, PIRSF and SUPERFAMILY and covers over 78% of all proteins in the Swiss-Prot and TREMBL components of UniProt (Mulder et al., 2005).

Computing our HP using InterPro clearly identified domains in seven out of eight HP and the eight's protein sequence is certainly too short to get a significant hit. All other databases failed to detect a domain and ProDom only showed a conserved region for the short protein Q96I26 (Table 3).

ProDom is a comprehensive database of protein domain families generated from global comparison of all available protein sequences. Recent improvements include the use of three-dimensional information from the SCOP database (Bru et al., 2005). It is complementing expert derived databases such as *PROSITE*, Pfam and SMART and is linked to InterPro.

Information on our HPs was only available for Q96GS8 identifying an ABC transporter domain (ProDom PD000006) probably limiting the use of ProDom to known proteins (Table 3).

Pfam is a large curated collection of protein multiple sequence alignments and profile hidden Markov models.¹ Predictions of non-domain regions are also included and contain active site residue mark up (Bateman et al., 2002). Pfam was useful for indication of domains in four out of eight HPs in our case (Table 3) and cannot be recommended as the only or first choice for searching of HPs.

SMART stands for Simple Modular Architecture Research Tool and is used for the identification and annotation of protein domains providing a platform for comparative studies of complex domain architecture (Letunic et al., 2004). Completed metazoan genomes are integrated and predictions of orthology are allowed.

As shown in Table 3, SMART was providing useful information on three HPs.

4.2. Motif analysis

Motif analysis is an obligatory step in the identification and characterisation of HPs. Detection of common motifs among proteins in particular with absent or low sequence identities (e.g. less than 30%) may provide important clues for function or classification of HPs into proper families (Rost and Valencia, 1996).

A series of signature databases are publically available and are often combined with sequence cluster and domain databases including *PROSITE*, PRINTS, Pfam, ProDom, Blocks, SMART and InterPro (see below).

A potent method for motif searches represents the use of ELM (eukaryotic linear motif server; Table 4), a resource for investigating candidates short non-globular functional and structural motifs/sites in eukaryotic proteins. Sequence comparisons with short motifs are difficult to evaluate because the usual significance assessments are inappropriate. Therefore, the server is implemented with several logical filters to eliminates false positives, that are for cell compartment, globular domain clash and taxonomic range. Short linear peptide motifs are used for cell compartment targeting, protein–protein interactions and regulation by post-translational modifications (Puntervoll et al., 2003).

Results are expressed as providing motif name, positions, ELM description, pattern and presence or absence of eukaryotic linear motifs. For each HP at least one linear motif was detectable and there was a wide overlap of motifs. 19 linear motifs were present in one protein exclusively.

Analysing HP studied herein showed typical results as shown in Table 4:

ELM analysis of HP FLJ20331 protein (Q9BTR7), e.g. revealed two signatures at positions 230–232 and 329–331

¹ Speech recognition uses probabilistic models to interpret a sequence of sounds and the same technique has been adapted to find domains in protein sequences of amino acids (Coin et al., 2003).

Table 3

Domains, motifs, repeats and protein families of hypothetical proteins in MB cell line (DAOY)

Accession number sequence, protein name	Subcellular localisation	Protein family/domain (database, ID number)	Motifs/repeats (database; ID number)
Q9BTR7 FLJ20331 protein [Fragment] Sequence length: 581 aa	Cytoplasm (score: 0.650) Non-secretory protein	Histidine acid phosphatase family InterPro: IPR000560 Histidine acid phosphatases active site signature (HIS_ACID_PHOSPHAT_2) PROSITE: PS00778	Leucine-rich repeat InterPro: IPR001611 PROSITE: PR00019 PFAM: PF00560 PROSITE: IPR003591 Typical subtype SMART: SM00369 PROSITE: PS50506 (PROFILE) Found in Polycystin cation channel Domain (IPR006228) Motifs: N-myristoylation site PROSITE: PS00008 N-glycosylation site PROSITE: PS00001 Protein kinase C phosphorylation site PROSITE: PS00005 Casein kinase II phosphorylation site PROSITE: PS00006 Tyrosine kinase phosphorylation site PROSITE: PS00007
Q81Y67 RAVER1 Sequence length: 606 aa	Nucleus (score: 0.964) Non-secretory protein	Domain: RNA-binding region RNP-1 (RNA recognition motif) InterPro: IPR000504 PFAM: PF00076 SMART: SM00360 PROSITE: PS50102 (Profile) Found in Paraneoplastic encephalomyelitis antigen family (IPR002343)	Motifs: Proline-rich region profile PROSITE: PS50099 Protein kinase C phosphorylation site PROSITE: PS00005 Casein kinase II phosphorylation site PROSITE: PS00006 N-myristoylation site PROSITE: PS00008 cAMP- and cGMP-dependent protein kinase phosphorylation site PROSITE: PS00004 Bipartite nuclear targeting sequence PROSITE: PS00015
Q96GS8 Hypothetical protein ABCF3 Sequence length: 709 aa	Nucleus (score: 0.760) Non-secretory protein	Domain: (1) ABC transporter related domain InterPro: IPR003439 ABC transporter PRODOM PD000006 PFAM PF00005 PROSITE PS00211 (Profile) (ABC_TRANSPORTER_1) PROSITE PS50100 (PROFILE) (DA_BOX) PROSITE PS50893 (PROFILE) (ABC_TRANSPORTER_2) (2) AAA ATPase InterPro: IPR008693 SMART: SM00382 Found in Antigen peptide transporter 2 family; IstB-like ATP-binding protein; Origin of replication binding protein family, . . .	Motifs: N-myristoylation site PROSITE: PS00008 Protein kinase C phosphorylation site PROSITE: PS00005 Casein kinase II phosphorylation site PROSITE: PS00006 cAMP- and cGMP-dependent protein kinase phosphorylation site PROSITE: PS00004 Amidation site PROSITE: PS00009 Tyrosine kinase phosphorylation site PROSITE: PS00007 ATP/GTP-binding site motif A (P-loop) PROSITE: PS00017

Table 3 (Continued)

Accession number sequence, protein name	Subcellular localisation	Protein family/domain (database, ID number)	Motifs/repeats (database; ID number)
Q9Y512 SAM50-like protein CGI-51 Sequence length: 469aa	Microbody (peroxisome) (score: 0.748) Non-secretory protein	Bacterial surface antigen (D15) family InterPro: IPR000184 PFAM: PF01103	Motifs: N-myristoylation site PROSITE: PS00008 Casein kinase II phosphorylation site PROSITE: PS00006 cAMP- and cGMP-dependent protein kinase phosphorylation site PROSITE: PS00004 N-glycosylation site PROSITE: PS00001 Protein kinase C phosphorylation site PROSITE: PS00005 Tyrosine kinase phosphorylation site PROSITE: PS00007 Tyrosine sulfation site PROSITE: PS00003
Q96I26 PTPN11 protein Sequence length: 50 aa	Nucleus (score: 0.650) Non-secretory protein	PRODOM PD498634	Motifs: cAMP- and cGMP-dependent protein kinase phosphorylation site PROSITE: PS00004 Casein kinase II phosphorylation site PROSITE: PS00006 Protein kinase C phosphorylation site PROSITE: PS00005
Q9P0J3 Putative 55 kDa protein Sequence length: 505 aa	Cytoplasm (score: 0.650) Non-secretory protein	Protein of unknown function UPF0027 family InterPro: IPR001233 PROSITE: PS01288	Motifs: Casein kinase II phosphorylation site PROSITE: PS00006 N-myristoylation site PROSITE: PS00008 Protein kinase C phosphorylation site PROSITE: PS00005 N-glycosylation site PROSITE: PS00001 cAMP- and cGMP-dependent protein kinase phosphorylation site PROSITE: PS00004 Tyrosine sulfation site PROSITE: PS00003
Q9JJG2 Mus musculus brain cDNA, clone MNCb-2622, similar to B033102 KIAA1276 protein Sequence length: 596 aa	Nucleous (score: 0.300) Non-secretory protein	Prefoldin family: Interpro: IPR009053 Superfamily: SSF46579 Domain: Ubiquitin interacting motif InterPro: IPR003903 PFAM: PF02809.8 Found in: Machado-Joseph disease protein MJD	Motifs: N-myristoylation site PROSITE: PS00008 Protein kinase C phosphorylation site PROSITE: PS00005 Casein kinase II phosphorylation site PROSITE: PS00006 cAMP- and cGMP-dependent protein kinase phosphorylation site PROSITE: PS00004 Amidation site PROSITE: PS00009 Tyrosine kinase phosphorylation site PROSITE: PS00007 ATP/GTP-binding site motif A (P-loop) PROSITE: PS00017
Q9BT99 Ras association (RalGDS/AF-6) domain family 5, isoform D Sequence length: 390 aa	Microbody (peroxisome) (score: 0.300) Non-secretory protein	Domains: (1) Ras associated domain InterPro: IPR000159 PFAM: PF00788 SMART: SM00314 PROSITE: PS50200	Motifs: Proline-rich region profile PROSITE: PS50099 Arginine-rich region profile PROSITE: PS50323 Casein kinase II phosphorylation site

Table 3 (Continued)

Accession number sequence, protein name	Subcellular localisation	Protein family/domain (database, ID number)	Motifs/repeats (database; ID number)
		(2) Protein kinase C, phorbol ester/diacylglycerol binding domain InterPro: IPR002219 PFAM: PF00130 SMART: SM00109 PROSITE: PS00479 (profile) PROSITE: PS50081 (profile)	PROSITE: PS00006 Protein kinase C phosphorylation site PROSITE: PS00005 <i>N</i> -myristoylation site PROSITE: PS00008 Tyrosine kinase phosphorylation site PROSITE: PS00007 cAMP- and cGMP-dependent protein kinase phosphorylation site PROSITE: PS00004

representing *N*-arginine dibasic convertase cleavage sites, and the pattern .RKIRR[^KR] (linear motif).

It is, however, intriguing to analyse the short HP Q96I26 that contained several ELMs including a specific ELM representing a part of the peroxisomal matrix import system (Table 4).

It would be important to investigate correlation of linear motifs with protein–protein interactions as 14-3-3 protein interaction motif 2 was observed in HP Q9P0J3 but protein–protein interaction databases do not list this interaction (see Table 5).

4.3. Functional analysis by protein–protein interaction and protein association databases

There is a series of protein–protein interaction and protein association databases. Here, we describe two widely used programmes, InterWeaver and STRING.

InterWeaver is discovering potential protein–protein interactions with online evidence automatically extracted from protein interaction databases, literature abstracts, domain fusion events and domain interactions. It searches online protein interaction databases DIP (Xenarios et al., 2002), BIND (Alfarano et al., 2005) and Protein Data Bank (Westbrook et al., 2003) for experimentally derived protein interactions and complexes. Two approaches are existing when new protein sequences are computed. The first, homology based approach finds similar proteins in other species and then searches protein interaction databases and biomedical literature to identify partners. In the domain-based approach, the system searches databases of domain-fusion events and putative domain interactions to propose reaction partners (Zhang and Ng, 2004). Computing our HPs into the programme (Table 5), e.g. revealed another hypothetical interacting protein Q9VIW7 from *Drosophila* with a domain of unknown function UPF0027, which does not take us any further. An interacting domain, Zinc finger, C2H2 type, was however predicted that may represent a first faint clue on pathway-involvement.

In HP Q9Y512 and Q96I26 *InterWeaver* searches provided no information on protein–protein interactions and therefore the application of *InterWeaver* did not add information to other search strategies.

STRING is a database of predicted functional associations between proteins (von Mering et al., 2003) and functional links. Protein–protein interactions are not limited to direct physical binding and proteins may also interact indirectly as e.g. by sharing a substrate, regulating each other transcriptionally, or participate in multi-protein assemblies. Non-physical interactions along with physical interactions form so-called genomic context or non-homology-based inference methods including STRING. Indigo, Clusters of Orthologous Group (COG; Tatusov et al., 2001), Predictome (Mellor et al., 2002) and SNAPer (Kolesov et al., 2002) only rely on a single form of genomic context and are not integrated. In contrast, STRING indicates reliability of predictions. Using STRING in the analysis of our HPs (Table 5) failed to show prediction of an association for Q9JJG2 and the short HP Q96I26. In all other cases significant associations (score > 0.4) (von Mering et al., 2005) to other proteins were detectable. For three proteins (Q9Y512, Q96I26 and Q9P0J3) where no functional domain was predicted in several databases, STRING showed an association in form of co-expression with known proteins in two cases and for protein Q9P0J3 at least association (neighbourhood) to a protein cluster of ortholog proteins (COG0430) was indicated (Table 5). We do not learn to much about our three abovementioned HPs using STRING, in particular as no physical interaction and only co-expression was observed.

5. Homology searches

A sequence can be recognised as a homolog of a known protein if the pair-wise sequence identity/similarity exceeds a statistically driven threshold (e.g. more than 30% sequence identity of an *E*-value less than 0.001 (Chothia and Lesk, 1986). This is, however, not a strict definition as homology is a qualitative statement—proteins are either homologous or not (Reeck et al., 1987) and the term can be used only in context with evolution, common ancestry. Homology is not expressed as percentage of sequence similarities or identities and homology does not necessarily mean comparable function. Homology searches are therefore not helpful for determination of protein function and are mainly forming a concept for protein family classification and studies on phylogenesis and evolution.

Table 4
ELM (eukaryotic linear motif)-database search results

Motif-name	ELM description	Pattern ^a	Q9BRT7	Q8IY67	Q96GS8	Q9Y512	Q96I26	Q9P0J3	Q9GJJ2	Q9BT99
CLV_NDR_NDR_1	N-Arg dibasic convertase (nardilysine) cleavage site	.RK RR[^KR	+	+	+	+	+	+	+	+
CLV_PCSK_FUR_1	Furin (PACE) cleavage site	R.[RK]R.	–	–	–	–	–	–	–	+
CLV_PCSK_PC1ET2_1	NEC1/NEC2 cleavage site	KR.	+	+	+	–	–	+	+	–
CLV_PCSK_PC7-1	Proprotein convertase 7 (PC7, PCSK7) cleavage site	[RK].[AILMFV][LTKF]	–	–	–	–	–	+	–	–
CLV_PCSK_SKI1_1	Subtilisin/Kexin isozyme-1 (SKI1) cleavage site	[RK].[AILMFV][LTKF].	+	+	+	+	–	+	+	+
LIG_14-3-3_2	14-3-3 proteins interaction motif 2	R.[SYFWTQAD].	–	–	–	–	–	+	–	–
LIG_14-3-3_3	Consensus derived from natural interactors which do not exactly match the mode 1 and mode2 ligands	[ST].[PLM] [RHK][STALV.][ST] .[PESRDIF]	–	–	+	+	–	+	–	–
LIG_AP2alpha_1	FxDxF motif; Responsible for the accessory endocytic proteins to the appendage; Of the alpha-subunit of adaptor protein; Complex AP-2	F.D.F	–	–	+	+	+	–	–	–
LIG_Clathr_ClatBox_1	Clathrin binding motif	L[IVLMF].[IVLMF][DE]	–	–	–	–	–	+	–	–
LIG_CYCLIN_1	Predicted protein should have the MOD_CDK site. Also used by cyclin inhibitors	[RK].L.{0,1}[FYLVMP]	+	+	+	+	–	+	+	+
LIG_FHA_1	FHA domain interaction; Motif 1, threonine phosphorylation is required	T.[ILA]	+	+	+	+	–	+	+	+
LIG_IQ	Calmodulin binding helical motif	...[SACLIVTM] ..[ILVMFCT] Q.{3}[RK].{4,5}[RKQ]..	–	+	–	–	–	–	–	–
LIG_NRBOX	Nuclear receptor box motif (LXXLL)	NLRLLLL	+	+	–	–	–	–	–	–
LIG_PP1	Is a conserved protein phosphatase 1 catalytic subunit)- binding motif	..[RK].{0,108VI}[^P][FW].	–	+	+	+	–	–	–	–
LIG_PDZ_3	Class III PDZ domains binding motif	.[DE].[IVL]	+	+	+	+	+	+	+	+
LIG_SH2-PTP2	SH-PTP2 and phospholipase C-gamma Src Homology 2 (SH2) Domains binding motif	Y[IV].[VILP]	–	–	–	–	–	–	+	–
LIG_CtBP	PxDLS motif that interacts with CtBP Protein	[PG][LVIPME][DENS] L[VASTRAGE]	–	+	–	–	–	–	–	–
LIG_RB	Interacts with Retinoblastoma protein	[LI].C.[DE]	–	–	+	–	–	–	–	–
LIG_SH2_GRB2	GRB2-like Src homology 2 (SH2) domains binding motif	Y.N.	–	–	–	–	–	+	–	–
LIG_SH2_SRC	Src-family Homology 2 (SH2) domains binding motif	Y[VLTFC].	–	–	–	–	–	+	–	–
LIG_SH2_STAT3	It binds STAT3 SH2 domain	Y.Q	–	–	+	–	–	–	–	–
LIG_SH2_STAT5	STAT5 src; Homology 2 domain binding motif	Y[VLTFC].	+	+	+	+	–	+	+	+

LIG_MYND	PxLxP motif, MYND ligand	...[SACLIVTM] ..[ILVMFCT] Q.{3}[RK]. {4,5}[RKQ]..	-	+	-	-	-	-	-	-
LIG_RGD	Found in proteins in extracellular matrix and recognized by different of integrin family	RGD	-	-	+	-	-	-	-	-
LIG_SH3_1	This motif recognized by class I SH3 domains	[RKY]..P..P	-	+	-	-	-	-	-	+
LIG_SH3_2	This is the motif recognized by class II SH3 domains	P..P.[KR]	-	+	-	-	-	-	+	+
LIG_SH3_3	Is the motif recognized By class I recognition specificity	..[PV]..P	+	+	-	+	-	+	+	+
LIG_SH3_4	This is the motif recognized by those SH3 domains with a non-canonical class II recognition specificity	KP..[QK]..	-	-	+	-	-	-	-	-
LIG_TRAF6	TRAF6 binding site	..P.E..[FYWHDE].	+	-	-	-	-	-	-	+
LIG_TRAF2_1	Member of the tumor necrosis factor superfamily	[PSAT].[QE]E	-	+	+	+	-	-	+	+
LIG_WW_1	PPXY is the motif recognized by WW domains of Group I	PP.Y	-	-	-	-	-	-	-	+
LIG_WW_3	WW domain of group III binding motif	.PPR.	-	+	-	-	-	-	-	+
LIG_WW_4	Class IV WW domains interaction motifs	..[ST]P.	-	+	+	+	-	+	+	+
MOD_CK1_1	CK1 phosphorylation site	S..([ST])..	+	+	+	+	-	+	+	+
MOD_CK2_1	CK2 phosphorylation site	...([ST])..E	+	+	+	+	+	+	+	+
MOD_Cter_Amidation	Peptide C-terminal amidation	(.)G[RK][RK]	-	-	-	-	-	-	+	-
MOD_GlcNHglycan	Glucosaminoglycan attachment site	[ED]{0,3}.(S)[GA]	+	+	+	+	+	+	+	-
MOD_GSK3_1	GSK3 phosphorylation recognition site	...([ST])...[ST]	+	+	+	+	+	+	+	+
MOD_N-GLC_1	Generic motif for N-glycosylation	.(N)[^P][ST]..	+	-	+	+	-	+	-	-
MOD_N-GLC-2	Atypical motif for N-glycosylation site	(N)[^P]C	-	-	-	+	-	+	-	+
MOD_NMyristoyl	Generic motif for N-Myristoylation	^MG ^G [^EDRKHPFYW] ..[STAGCN][^P]	-	-	-	+	-	-	-	-
MOD_PKA_1	Protein kinase (signaling)	[RK][RK].[ST]....	-	-	+	+	+	+	-	+
MOD_PK_1	Phosphorylase kinase phosphorylation site	[RK]..(S)[VI]..	-	-	-	+	-	-	-	-
MOD_PKA_2	PKA phosphorylation site	.R.([ST])...	+	+	+	+	+	+	-	+
MOD_PLK	Site phosphorylated by the Polo-like-kinase	[DE].[ST] [ILFWMVA]..	+	+	+	+	+	+	+	+
MOD_ProDKin_1	Proline-directed kinase (MAPK) phosphorylation	...([ST])P..	-	+	+	+	-	+	+	+
MOD_SUMO	Motif recognized for modification by SUMO-1	[DER]. . .L[LV]	-	-	+	+	-	+	+	-

Table 4 (Continued)

Motif-name	ELM description	Pattern ^a	Q9BRT7	Q8IY67	Q96GSS	Q9Y512	Q96I26	Q9P0J3	Q9GJ12	Q9BT99
TRG_ENDOCYTIC_2	Tyrosine-based sorting signal for interaction with Adaptor protein mu	Y..[LMVIF]	-	-	+	+	-	+	-	+
TRG_LysEnd_ApsAcLL_1	Sorting and internalisation signal found in the cytoplasmic Juxta-membrane region of type I transmembrane proteins	[DER]...L[LVII]	-	+	-	+	-	+	-	+
TRG_WXXXXY/F	Specific ELM; Present in Pex5p and binding to Pex13p and Pex14p. Part of the peroxisomal matrix protein import system	W...[FY]	-	-	-	-	+	-	-	-

^a Pattern description can be found in <http://elm.eu.org/help.html>.

6. Protein characterisation by physicochemical properties

6.1. Amino acid composition and protein stability

Amino acid compositional analysis has been shown to be informative in a number of studies. Composition is related to biological characteristics of a protein such as its location (intracellular or extracellular), function (enzymes or non-enzymes), structural features, the presence of disulfide bonds and the folding type, to name a few (Nishikawa et al., 1983; Nakashima and Nishikawa, 1992, 1994).

Peptide chains could be assigned as either cytoplasmic or extracellular, solely from the analysis of sequence composition (Nakashima and Nishikawa, 1994).

Compositional differences between cytoplasmic and secretory proteins have been used to develop software for predicting transmembrane helices in integral membrane proteins that in turn use analysis of charge bias and hydrophobicity (Guruprasad et al., 1990; von Heijne, 1992).

In Table 6, amino acid composition of HP is provided and the percentage of individual amino acids is given. For protein Q9BTR7 the percentage of leucine is 18.8% and this correlates with the 14 leucine-rich repeats (Table 2). High leucine content is however not representative for leucine-rich repeats as clearly seen for protein Q8IY67 (Tables 3 and 6). The total number of negatively or positively charged amino acids is listed and fairly identifies basic or acidic proteins as in the case of Q9JG2 (Table 6).

Biological characteristics of proteins such as in vivo stability and structural features have been found to correlate with distinct patterns of amino acid composition. It was observed that the occurrence of certain dipeptides was significantly different in the unstable proteins compared to the stable ones (Guruprasad et al., 1990; Reddy, 1996). Low complexity protein sequences with lower proportion of distinct dipeptides have non-globular shapes compared to sequences of high complexity (Nandi et al., 2003).

6.2. Hydrophobicity

Average hydrophobicity and charge are well conserved during evolution (for review: Saraf et al., 2004).

Kauzmann (1959) described important features of the thermodynamic stabilities of proteins. The hydrophobic effect is recognized as an important contributor to the stability of proteins and an important determinant of their structural patterns (Lesk, 2003).

The concept of hydrophobicity has been addressed by researchers in all aspects of science, particularly in the field of biology and chemistry. Over the past several decades, the study of the hydrophobicity of biomolecules, particularly amino acids has resulted in the development of a variety of hydrophobicity scales. Until now different hydrophobic scales are known (Kyte and Doolittle, 1982; Bastolla et al., 2005).

Table 5
Prediction of protein–protein interactions of identified hypothetical proteins in MB cell line (DAOY)

Accession number	Protein name	Gene name	Identification of potential protein–protein interaction (InterWeaver database)						Prediction of functional associations (STRING database) (text mining; coexpression; experiment; database; neighbourhood)					
			By homologs			By domains			Accession number	Protein name	Score	Active prediction methods		
			Species	Number of protein homologs	Number of interacting Proteins	Name of potential Interactors	Evidence from experiment (Database, literature)	Homolog evalue						
Q9BTR7	FLJ20331 protein [Fragment]	FLJ20331	Fruit fly	270	495	NP_476712.1 Ran GTPase-activating protein (RanGAP)	DIP ^a ; BIND ^b	CG9031-PA (NP_609665) (e-value: 2.8e-14)	–	COG0515	Serine/threonine protein kinase	0.981	Text mining; coexpression; experiment; database; neighbourhood	
			<i>Caenorhabditis elegans</i>	35	113	Putative protein, with 5 coiled coil domains, of bilateral origin (NP_501530)	DIP	Scribbled, LEThal LET-413 (75.3 kD) (e-value: 3.9e-24)						
			Baker's yeast	60	276	Protein kinase DBF2 (S64387)	PDB ^c	Adenylate cyclase (OYBY) (e-value: 5.0e-16)						
			Human	24	24	Platelet glycoprotein Ib alpha chain [Precursor] (P07359)	PDB	Platelet glycoprotein Ib alpha chain [Precursor] (P07359)						
			Bacterium monocytogenes	7	7	Epithelial-cadherin [Precursor] (P12830)	PDB	Epithelial-cadherin [Precursor] (P12830) (e-value: 3.2e-07)						
		Human	–	–	Leucine-rich repeat protein SHOC-2 (Q9UQ13) (e-value: 5.9e-34)	Biomedical literature	–							
Q8IY67	RAVER1	RAVER1	Human	17	17	Polyadenylate-binding protein 1 (P11940)	PDB	Polyadenylate-binding protein 1 (P11940) (e-value: 0.004)	2 RRM (RNA recognition motif) by fusion domain: 19 interactings domains	P09327	Villin 1	0.603	Text mining	
			Human	–	–	Splicing factor, proline-and glutamine-rich (P23246) (e-value: 7.8e-13)	biomedical literature	–		COG5354	Uncharacterized protein, contains Trp-Asp (WD) repeat	0.998	Gene fusion, coexpression, experiments, database, text mining	
			Fruit fly	132	433	Score: 14 NP_525033 (CG4262-PA) (Elav protein (Embryonic lethal abnormal visual protein) (P23241) (94% identity)	DIP; BIND	NP_572842 (CG4396-PA) (e-value: 1.5e-07)	by domain-domain interaction: 683 interacting domains					
			<i>Caenorhabditis elegans</i>	5	40	NP_492508 (MEChanosensory abnormality MEC-8) [mec-8] (Q22039)	DIP	NP_496057 (EXCretory canal abnormal EXC-, ELAV type RNA binding protein) (e-value: 5.7e-10)						
			Baker's yeast	37	210	Eukaryotic initiation factor 4F subunit p130 (P39936) (87% identity)	DIP	Polyadenylate-binding protein (DNBYPA) (e-value: 2.2e-06)						
			Mouse	2	2	ELAV-like protein 3 (Q60900)	PDB	ELAV-like protein 3 (Q60900) (e-value: 2.5e-07)						
Q96GS8	Hypothetical protein ABCF3	ABCF3	Fruit fly	53	81	CG17293-PA (NP_609217)	DIP; BIND	CG9330-PA (NP_649129) (e-value: 0)	by domain fusion: 11 interaction domains	COG0086	DNA-directed	0.999	Database text mining, experiment	
			Baker's Yeast	74	176	Replication factor C chain RFC2 (S45531)	DIP	CG9330-PA (NP_649129) (8.0e-27)	by domain-domain interaction: 149 interacting domains		RNA polymerase, beta subunit/160 kDa subunit			
			<i>Caenorhabditis elegans</i>	8	12	Nuclear Hormone Receptor (53.3 kD)	DIP	1 inhibitor (69.0 kD) (NP_499717) (e-value: 3.2 e-07)						
			<i>Helicobacter pylori</i> (strain 26695)	15	20	Phosphoserine aminotransferase (serC) (NP_207530)	BIND; DIP	cell division protein (NP_207541) (e-value: 2.7e-06)						

Table 5 (Continued)

Accession number	Protein name	Gene name	Identification of potential protein–protein interaction (InterWeaver database)						Prediction of functional associations (STRING database) (text mining; coexpression; experiment; database; neighbourhood)				
			By homologs			By domains			Accession number	Protein name	Score	Active prediction methods	
			Species	Number of protein homologs	Number of interacting Proteins	Name of potential Interactors	Evidence from experiment (Database, literature)	Homolog evalue					
			“Bacillus coli” migula 1895	6	11	Vitamin B12 import system permease protein btuC (P06609)	PDB	Vitamin B12 import system permease protein btuC (P06609) (e-value: 3.1e-04)					
			Human	10	7	ATP-binding cassette, sub-family D, member 2 (NP_005155)	BIND	ATP-binding cassette, sub-family D, member 2 (NP_005155) (e-value: 0.001)					
			Mouse	1	1	ATP-binding cassette, sub-family D, member 2 (NP_005155)	BIND	ATP-binding cassette, sub-family D, member 2 (4.3e-04)					
			<i>Methanococcus janaschii</i>	3	3	ABC transporter ATP-binding protein MJ0796 (Q58206)	PDB	ABC transporter ATP-binding protein MJ0796 (Q58206) (e-value: 5.7e-06)					
Q9Y512	SAM50-like protein CGI-51	–	–	–	–	–	–	–	Interacting domain: 2Fe-2S iron-sulfur cluster binding domain	COG0764	D-3-hydroxydecanoyl-(acyl carrier-protein) dehydratase	0.907	Gene fusion, neighbourhood
Q96I26	PTPN11 protein	–	–	–	–	–	–	–	–	–	–	–	–
Q9P0J3	Putative 55 kDa protein	–	Fruit fly	2	2	CG2052-PB (NP_726568)	DIP	CG9987-PA (NP_609965) (e-value: 0)	interacting domain: Zinc finger, C2H2 type	COG0430	RNA 3-terminal phosphate cyclase	0.826	Text mining, neighbourhood coexpression
Q9JJG2	Mus musculus brain cDNA, clone MNCh-2622, similar to AB033102 KIAA1276 protein	AB041544	Baker's yeast	188	383	Protein kinase (TVBY8)	DIP	Myosin-like protein 1 (CAA82174) e value: 4.3e-10)	–	–	–	–	–
			Fruit fly	800	1541	CG3610-PA (NP_650396)	DIP; BIND	CG6450-PC (NP_525064) (e-value: 1.4e-13)					
			<i>Caenorhabditis elegans</i>	77	184	Putative nuclear protein family member, nematode specific, GEX (NP_497245)	DIP	Non-muscle myosin (nmy-1) (NP_508504) (e-value: 7.8e-13)					
			Human	27	52	Keratin 5, type II, epidermal (A29904)	DIP	Desmoplakin I (A38194) (e-value: 1.1e-10)					
			Mouse	6	11	Desmoplakin I (A38194)	DIP	Desmin (A54104) (e-value: 2.9e-06)					
			Rat	1	2	Neurofilament triplet L protein (A21762)	DIP	Neurofilament triplet H protein (A37221) (e-value: 0.001)					
			<i>Helicobacter pylori</i> (strain 26695)	3	4	Flagellar basal-body rod protein (FlgB) (Proximal rod protein) (NP_208350)	BIND	Hypothetical protein HP1143 (NP_207934) (e-value: 0.002)					
			European rabbit	2	2	Tropomyosin 1 alpha chain (P58772)	PDB	Tropomyosin 1 alpha chain (P58772) (e-value: 0.003)					
			Sus scrofa	4	4	Tropomyosin 1 alpha chain (P42639)	PDB	Tropomyosin 1 alpha chain (P42639) (e-value: 0.004)					
Q9BT99	Ras association (RalGDS/AF-6) domain family 5, isoform D	RASSF5	Fruit fly	6	9	CG13287-PA (NP_647994)	DIP; BIND	CG4656-PA (NP_651126) (e-value: 0.005)	–	ENSP00000336616	Ras association (RalGDS/AF-6) domain family 3	0.730	Homology text mining
			Human	–	–	Ras association domain family 2 (P50749) (e-value: 1.5e-08)	Biomedical literature	–	–	KOG1613	Exosomal 3-5 exoribonuclease complex, subunit Rrp	0.859	Experiments

^a Database of Interacting Proteins.

^b Biomolecular Interaction Network Database.

^c Protein Data Bank.

Table 6
Results of physiochemical properties of hypothetical proteins in MB cell line (DAOY)

Accession number	Protein name	Amino acid composition						Instability index	Aliphatic index	GRAVY ^a				
Q9BTR7	FLJ20331 protein [Fragment]	Ala (A)	20	3.4%	Leu (L)	109	18.8%	47.53 (unstable)	114.42	-0.192				
		Arg (R)	31	5.3%	Lys (K)	33	5.7%							
		Asn (N)	48	8.3%	Met (M)	12	2.1%							
		Asp (D)	29	5.0%	Phe (F)	16	2.8%							
		Cys (C)	9	1.5%	Pro (P)	27	4.6%							
		Gln (Q)	21	3.6%	Ser (S)	52	9.0%							
		Glu (E)	46	7.9%	Thr (T)	29	5.0%							
		Gly (G)	16	2.8%	Trp (W)	3	0.5%							
		His (H)	10	1.7%	Tyr (Y)	7	1.2%							
		Ile (I)	37	6.4%	Val (V)	26	4.5%							
		Asx (B)	0	0.0%										
		Glx (Z)	0	0.0%										
		Xaa (X)	0	0.0%										
		Total number of negatively charged residues (Asp + Glu): 75												
		Total number of positively charged residues (Arg + Lys): 64												
Q8IY67	RAVER1	Ala (A)	69	11.4%	Met (M)	7	1.2%	51.67 (unstable)	85.15	-0.276				
		Arg (R)	37	6.1%	Asn (N)	15	2.5%							
		Asp (D)	19	3.1%	Cys (C)	11	1.8%							
		Gln (Q)	33	5.4%	Glu (E)	32	5.3%							
		Gly (G)	64	10.6%	His (H)	14	2.3%							
		Ile (I)	8	1.3%	Leu (L)	91	15.0%							
		Lys (K)	21	3.5%	Phe (F)	14	2.3%							
		Pro (P)	67	11.1%	Ser (S)	43	7.1%							
		Thr (T)	26	4.3%	Trp (W)	4	0.7%							
		Tyr (Y)	10	1.7%	Val (V)	21	3.5%							
		Asx (B)	0	0.0%										
		Glx (Z)	0	0.0%										
		Xaa (X)	0	0.0%										
		Total number of negatively charged residues (Asp + Glu): 51												
		Total number of positively charged residues (Arg + Lys): 58												
Q96GS8	Hypothetical protein ABCF3	Ala (A)	58	8.2%	Leu (L)	88	12.4%	45.14 (unstable)	91.35	-0.386				
		Arg (R)	57	8.0%	Lys (K)	35	4.9%							
		Asn (N)	23	3.2%	Met (M)	13	1.8%							
		Asp (D)	40	5.6%	Phe (F)	27	3.8%							
		Cys (C)	7	1.0%	Pro (P)	26	3.7%							
		Gln (Q)	37	5.2%	Ser (S)	49	6.9%							
		Glu (E)	62	8.7%	Thr (T)	25	3.5%							
		Gly (G)	49	6.9%	Trp (W)	5	0.7%							
		His (H)	15	2.1%	Tyr (Y)	18	2.5%							
		Ile (I)	29	4.1%	Val (V)	46	6.5%							
		Asx (B)	0	0.0%										
		Glx (Z)	0	0.0%										
		Xaa (X)	0	0.0%										
		Total number of negatively charged residues (Asp + Glu): 102												

Table 6 (Continued)

Accession number	Protein name	Amino acid composition						Instability index	Aliphatic index	GRAVY ^a
		Total number of positively charged residues (Arg + Lys): 92								
Q9Y512	SAM50-like protein CGI-51	Ala (A)	23	4.9%	Met (M)	9	1.9%	31.22 (stable)	83.11	−0.213
		Arg (R)	29	6.2%	Phe (F)	29	6.2%			
		Asn (N)	19	4.1%	Pro (P)	20	4.3%			
		Asp (D)	21	4.5%	Ser (S)	36	7.7%			
		Cys (C)	7	1.5%	Thr (T)	20	4.3%			
		Gln (Q)	13	2.8%	Trp (W)	7	1.5%			
		Glu (E)	33	7.0%	Tyr (Y)	13	2.8%			
		Gly (G)	55	11.7%	Val (V)	31	6.6%			
		His (H)	11	2.3%	Ile (I)	22	4.7%			
		Leu (L)	49	10.4%	Lys (K)	22	4.7%			
		Asx (B)	0	0.0%						
		Glx (Z)	0	0.0%						
		Xaa (X)	0	0.0%						
		Total number of negatively charged residues (Asp + Glu): 54								
		Total number of positively charged residues (Arg + Lys): 51								
Q96126	PTPN11 protein	Ala (A)	2	4.0%	Met (M)	4	8.0%	20.11 (stable)	46.80	−0.840
		Arg (R)	5	10.0%	Phe (F)	4	8.0%			
		Asn (N)	0	0.0%	Pro (P)	0	0.0%			
		Asp (D)	2	4.0%	Ser (S)	2	4.0%			
		Cys (C)	1	2.0%	Thr (T)	1	2.0%			
		Gln (Q)	1	2.0%	Trp (W)	1	2.0%			
		Tyr (Y)	0	0.0%	Glu (E)	6	12.0%			
		Gly (G)	7	14.0%	Val (V)	2	4.0%			
		His (H)	1	2.0%	Ile (I)	0	0.0%			
		Leu (L)	4	8.0%	Lys (K)	7	14.0%			
		Asx (B)	0	0.0%						
		Glx (Z)	0	0.0%						
		Xaa (X)	0	0.0%						
		Total number of negatively charged residues (Asp + Glu): 8								
		Total number of positively charged residues (Arg + Lys): 12								
Q9P0J3	Putative 55 kDa protein	Ala (A)	48	9.5%	Ile (I)	26	5.1%	40.51 (unstable)	85.54	−0.209
		Arg (R)	26	5.1%	Leu (L)	39	7.7%			
		Asn (N)	23	4.6%	Lys (K)	32	6.3%			
		Asp (D)	31	6.1%	Met (M)	19	3.8%			
		Cys (C)	9	1.8%	Phe (F)	16	3.2%			
		Gln (Q)	19	3.8%	Pro (P)	20	4.0%			
		Glu (E)	28	5.5%	Ser (S)	21	4.2%			
		Gly (G)	51	10.1%	Thr (T)	20	4.0%			
		His (H)	16	3.2%	Trp (W)	3	0.6%			
		Tyr (Y)	13	2.6%	Val (V)	45	8.9%			
		Asx (B)	0	0.0%						
		Glx (Z)	0	0.0%						
		Xaa (X)	0	0.0%						
		Total number of negatively charged residues (Asp + Glu): 59								
		Total number of positively charged residues (Arg + Lys): 58								

Q9JJG2	Mus musculus brain cDNA, clone MNCb-2622, similar to AB033102 KIAA1276 protein	Ala (A)	41	6.9%	Ile (I)	14	2.3%	56.29 (unstable)	69.26	-1.092					
		Arg (R)	43	7.2%	Leu (L)	62	10.4%								
		Asn (N)	11	1.8%	Lys (K)	57	9.6%								
		Asp (D)	29	4.9%	Met (M)	14	2.3%								
		Cys (C)	12	2.0%	Phe (F)	4	0.7%								
		Gln (Q)	54	9.1%	Pro (P)	17	2.9%								
		Glu (E)	87	14.6%	Ser (S)	43	7.2%								
		Gly (G)	23	3.9%	Thr (T)	31	5.2%								
		His (H)	13	2.2%	Trp (W)	6	1.0%								
		Tyr (Y)	9	1.5%	Val (V)	26	4.4%								
		Asx (B)	0	0.0%											
		Glx (Z)	0	0.0%											
		Xaa (X)	0	0.0%											
		Total number of negatively charged residues (Asp + Glu): 116													
		Total number of positively charged residues (Arg + Lys): 100													
		Q9BT99	RASSF5	Ala (A)	23	5.9%	Leu (L)				44	11.3%	66.06 (unstable)	81.77	-0.502
				Arg (R)	37	9.5%	Lys (K)				18	4.6%			
Asn (N)	9			2.3%	Met (M)	3	0.8%								
Asp (D)	17			4.4%	Phe (F)	13	3.3%								
Cys (C)	12			3.1%	Pro (P)	43	11.0%								
Gln (Q)	22			5.6%	Ser (S)	27	6.9%								
Glu (E)	25			6.4%	Thr (T)	20	5.1%								
Gly (G)	22			5.6%	Trp (W)	3	0.8%								
His (H)	5			1.3%	Tyr (Y)	10	2.6%								
Ile (I)	17			4.4%	Val (V)	20	5.1%								
Asx (B)	0			0.0%											
Glx (Z)	0			0.0%											
Xaa (X)	0			0.0%											
Total number of negatively charged residues (Asp + Glu): 42															
Total number of positively charged residues (Arg + Lys): 55															

^a Grand average of hydropathy.

6.2.1. ProtParam tool

ProtParam (<http://www.expasy.org/tools/protparam.html>) computes various physicochemical properties that can be deduced from a protein sequence. The parameters computed by ProtParam include the molecular weight, theoretical pI, amino acid composition, atomic composition, extinction coefficient, estimated half-life, instability index (estimate of the stability of a protein in a test tube), aliphatic index (relative volume occupied by aliphatic side chains), and grand average of hydropathy (GRAVY). The grand average of hydropathy (GRAVY) value for a peptide or protein is calculated as the sum of hydropathy (Kyte and Doolittle, 1982) values of all the hydrophobic amino acids, divided by the number of amino acid residues in the protein sequence. The GRAVY index above zero is for hydrophobic proteins (Kyte and Doolittle, 1982) and reflect hydrophobicity of the whole protein.

As expected no hydrophobic HP were detected due to the hydrophilicity of the analysis system (negative values in Table 6).

The aliphatic index is defined as the relative volume occupied by aliphatic side chains (alanine, valine, isoleucine, and leucine) and may be regarded as a positive factor for the increased thermostability of globular proteins (Ikai, 1980). And indeed the leucine rich protein Q9BTR7 is paralleling a high aliphatic index (Table 6).

The stability index provides an estimate of the stability of a protein in a test tube. Statistical analysis of 12 unstable and 32 stable protein has revealed (Guruprasad et al., 1990) that there are certain dipeptides, the occurrence of which is significantly different in unstable proteins compared with those in the stable ones. An instability index which is smaller than 40 predicts proteins as stable and above 40 as unstable. As shown in Table 6 two HP described herein are considered stable (Q9Y512; Q96I26).

7. Prediction of subcellular localisation

In most protein lists subcellular localisation (SL) is provided for known proteins in general and in particular for HPs. The close correlation between protein function and SLs is reported (Chou, 2000a). Prediction of SL may be relevant for inferences for possible function, annotation of genomes and designing proteomics experiments and characterising pharmacological targets.

Different categories of prediction methods for SL are by amino acid composition, known targeting sequences, sequence homology/or motifs and a combination of the first three categories (hybrid method: Chou and Cai, 2003). During the last decade several automated programs were developed to predict SL (Chou, 2000a,b; Chou and Elrod, 1999a,b; Chou and Cai, 2002, 2003, 2005a,b; Pan et al., 2003; Zhou and Doctor, 2003; Xiao et al., 2005). These program prediction algorithms generally consist of two cores: one is giving a mathematical expression to effectively represent a protein and the other is finding an operational equation for predicting SL effectively. The process of expressing a protein from the classical amino acid composition vector (Nakashima et al., 1986) to the pseudo amino

acid composition vector (Chou, 2001a; Pan et al., 2003; Zhou and Doctor, 2003; Xiao et al., 2005), to “quasi-sequence-order effect” (Chou, 2000a) and to the functional domain approach (Chou and Cai, 2002) reflects the development of defining a protein according to different mathematical representations. The “pseudo-amino acid composition” method includes sequence order effects and so improves quality of amino acid composition-based prediction of SL (Chou, 2001a). Further prediction progressing is developing “pseudo amino acid composition”-based complexity measure factors for determination of SL (Xiao et al., 2005). Additionally, by hybridising gene ontology and “pseudo amino acid composition” approaches, Chou and Cai (2005a) introduced a new method to predict the SL of proteins with multiplex location features.

The prediction process using simple geometry distance algorithm (Nakashima et al., 1986), Mahalanobis distance algorithm (Cedano et al., 1997), covariant discriminant algorithm (Chou and Elrod, 1999a,b; Chou, 2000a), neural networks (Reinhardt and Hubbard, 1998; Hua and Sun, 2001; Garg et al., 2005) and current support vector machine (SVM) algorithm (Chou and Cai, 2002) reflects the development of computation by means of different mathematical operations.

For prediction in the eukaryotic system PSORT II is considered the standard reference method (Nakai and Kanehisa, 1992; Nakai and Horton, 1999). The authors have been constructing a knowledge base by organising various experimental and computational observations as a collection of “if-then” rules, an expert system, which utilises this knowledge base, for predicting localisation sites of proteins only from the information on the amino acid sequence and the source origin.

Results from PSORT II analyses from our own dataset of HPs are presented in Table 3 and are expressed as subcellular site and *certainty* (score).

8. Signal peptide prediction

What is the significance of signal peptide prediction? The need to identify and predict SPs comes from the need to find more effective vehicles for the production of recombinant proteins (Nielsen et al., 1997; Chou, 2002) and of course knowledge on the SP is an essential part of defining and characterising a protein per se.

A SP comprises the N-terminal part of the amino acid chain and is cleaved off during translocation. The common structure is a positively charged n-region, followed by a hydrophobic h-region and a neutral but polar c-region. The (-3,-1) rule claims that residues at positions -3 and -1 relative to the cleavage site have to be small and neutral for correct cleavage (von Heijne, 1984).

A SP controls the entry of proteins to the secretory pathway in eukaryotes and prokaryotes. The extreme variation in length and sequence makes it difficult to develop a general algorithm to predict the SPs. Different methods like scaled window based on the HMM for SP prediction (Chou, 2001b) and most existing methods mainly based on neural networks (Nielsen et al., 1997; Claros et al., 1997) are used to predict SPs. Some new algorithms, coupling the probability model (Chou, 2001b) and

the subsite coupling principle with SVM (Wang et al., 2005a) were developed to overcome disadvantages of the neural network algorithm (NNA). These methods are faster and more accurate for prediction of SPs than NNA.

SignalP 3.0 Server, is an useful program for predicting the presence and location of SP cleavage sites in amino acid sequences from different organisms. The method incorporates a prediction of cleavage sites and signal peptide/non-signal peptide prediction based on a combination of several artificial neural networks and hidden Markov models (Bendtsen et al., 2004).

Results are expressed by listing either non-secretory protein or SP providing SP probability, maximal cleavage site probability and a graph. Using HPs studied herein with SignalP revealed that only non-secretory proteins were detected that may be due to incompleteness of sequences (Table 3).

9. Membrane proteins prediction

Membrane proteins in the strict sense are: integral membrane proteins containing at least one transmembrane domain (TMD). Different locations of proteins usually means different biological functions. Therefore, determination of function for new membrane proteins can be expedited significantly if we can find an effective algorithm for predicting their types and subcellular localisations (Chou and Elrod, 1999b; Chou and Cai, 2005b). Using amino acid composition will miss all sequence-order and sequence-length effects. In order to solve this problem, Chou (2001a) developed a new concept, “pseudo-amino acid composition” (Wang et al., 2004). The following progress was developing the supervised locally linear embedding (SLLE) technique which can reduce the operational space by extracting the essential features from the high-dimensional pseudo amino acid composition space, and the cluster-tolerant capacity can be increased accordingly (Wang et al., 2005b).

Predicting the type of a membrane protein from its primary sequence, or even just identifying whether the uncharacterized protein belongs to membrane proteins or not, is an important problem in bioinformatics and proteomics. To solve this problem, the GO-PseAA predictor was introduced and it is a hybrid system which combines gene ontology and pseudo amino acid composition. A dataset was constructed that contains both non-membrane and membrane proteins classified into five different types. To avoid redundancy and bias, none of the proteins included has > or =40% sequence identity to any other. The high success rates suggest that the GO-PseAA predictor can catch the core feature of the statistical samples concerned and may become an automated high throughput tool in molecular and cell biology (Chou and Cai, 2005c).

Different computer programs were developed for predicting the presence of transmembrane domains in proteins (review in Ahram and Springer, 2004).

9.1. TMHMM, HMMTOP and PHD programs

TMHMM (Krogh et al., 2001) is a membrane protein topology prediction method, which is based on a hidden

Markov model (HMM). This program is not only able to predict the transmembrane helices but also can discriminate between soluble and membrane proteins. A TMHMM prediction service is available at <http://www.cbs.dtu.dk/services/TMHMM/>. By using TMHMM server, we failed to get any evidence for transmembrane helices in our HPs.

HMMTOP (Tusnady and Simon, 2001) is another method based on HMM and is used for topology prediction of helical transmembrane proteins. This method is based on the hypothesis that the localisation of the transmembrane segments and the topology are determined by difference in amino acid distributions in various structural parts of these proteins rather than by specific amino acid compositions of these parts. By using this method, we could identify one transmembrane helix in HP Q9BT99 (122–139 sequence position). For the other HPs no transmembrane helices could be identified.

Both TMHMM and HMMTOP are global approaches. In opposite, PHD is a local approach and predicts one dimensional protein structure by profile-based neural networks (Rost, 1996). By using this program we identified helical transmembrane regions in two HPs. In HP Q9Y512 two and in HP Q9BT99 one transmembrane region were identified each

10. Structural bioinformatics

Structural bioinformatics is the subdiscipline of bioinformatics that focuses on representation, storage, retrieval, analysis and display of structural information at the atomic and subcellular spatial scales.

It is characterised by two goals: the creation of general purpose methods for manipulating information about biological macromolecules and the application of these methods to solving problems in biology and creating new knowledge.

The importance of structural bioinformatics is manifold: (1) creating an infrastructure for building up structural models from component parts, (2) gaining the ability to understand the design principles of proteins so that new functionalities can be created, (3) learning how to design drugs efficiently based on structural knowledge of their target, and (4) catalysing the development of simulation models that can give insight into function based on structural informations (Altman and Dugan, 2003).

The correct function of proteins depends on the very special and individual way of their folding, which is reflected by their 3D structure formed by the correct secondary, tertiary and quaternary structures.

The advantages of three-dimensional (3D) structures over sequence are in two distinct areas. Firstly, 3D structural information can provide new insights into biology by elucidating relatedness to proteins of known biological function that cannot be reached by sequence analysis alone. The protein structure is conserved over evolutionary time and therefore provides the opportunity to recognise homology that is undetectable by sequence comparison. Secondly, structural information can identify binding motifs and catalytic centers—even for proteins without a known biological function (Shapiro and Harris, 2000).

On the other side, computer-based 3D structure offers some advantages over experimental characterisation: they are faster and less expensive. Depending on the extent of post-translational modification and whether the identified proteins are multidomain structures which may need dissecting into domains or complexing with nucleic acids, proteins or polysaccharides before they can be crystallised and used for nuclear magnetic resonance (NMR) analysis, experimental 3D structure characterization will take more time and is more expensive.

Different concepts such as protein family, fold, and superfamily have been introduced (Hubbard et al., 1999; Orengo et al., 1999) and recently developed detailed taxonomies make the complex three-dimensional shapes of proteins easier to understand (Goldsmith-Fischman and Honig, 2003). Classification of protein structure in homologous families, superfamilies and fold is done by using different databases such as SCOP (Andreeva et al., 2004), CATH (Protein Structure Classification) (Pearl et al., 2005), FSSP (Database of Families of Structurally Similar Proteins) (Holm and Sander, 1996), CAMPASS (Cambridge database of Protein Alignments organised as Structural Superfamilies) (Sowdhmini et al., 1998) and HOMSTRAD (Homologous Structure Alignment Database) (Mizuguchi et al., 1998).

The Structural Classification of Proteins (SCOP) database is a comprehensive ordering of all proteins of known structure according to their evolutionary and structural relationships. The protein domains are fundamental units in SCOP database classification and are classified hierarchically into families, superfamilies, fold and classes, whose meaning has been discussed before (Murzin et al., 1995; Brenner et al., 1996). The first official SCOP release 10 years ago comprised 3179 protein domains grouped into 498 families, 366 superfamilies and 279-folds (Murzin et al., 1995). The seven main classes in the latest release (1.65) contain 40,452 domains organised into 2327 families and 800-folds. These domains correspond to 20619 entries in Protein Data Bank (PDB) (Westbrook et al., 2002). Statistics of the current and previous releases, summaries and full histories of changes and other information are available from the SCOP website (<http://scop.mrc-lmb.cam.ac.uk/scop/>) together with files encoding all SCOP data (Lo Conte et al., 2002). The sequences and structures of SCOP domains are available from the ASTRAL compendium (Chandonia et al., 2002) and HMMs of SCOP domains are available from the SUPERFAMILY database (Gough et al., 2001). As part of this project starting with release 1.63, Andreeva et al. (2004) initiated a refinement of the SCOP classification, which introduces a number of changes mostly at the levels below superfamily.

The main application of structural bioinformatics (Chou, 2004e) focused on prediction of three-dimensional structure of proteins and structure-function relationship (Chou, 2004e). Predicted 3D structure of many important proteins are already reported (Chou, 2004a,b,c,d, 2005; Chou et al., 1997, 1998, 1999, 2000, 2003; Chou and Howe, 2002; Du et al., 2004, 2005; Sirois et al., 2004).

The major 3D structure prediction categories are homology modeling (based on high sequence homology to a known

structure) (Sanchez and Sali, 1997), threading (based on remote sequence homology) (Marchler-Bauer and Bryant, 1997), and ab initio prediction (based on no-detectable homology) (Osguthorpe, 2000).

Since tertiary structure is better conserved than primary structure—if such a homology can be detected from sequence analysis—it is almost certain that the two proteins will share the same fold (Brenner, 2000).

10.1. Homology modeling

Homology modeling (Mosimann et al., 1995) is the most powerful method for determining the approximate structure of a protein with sequence similarity to known structures.

One out of seven of newly determined sequences has a similar sequence with known structure (Bork et al., 1992). Homology modeling contains six different steps: (1) finding a template protein with known structure for target protein (normally from the same family), (2) using different alignment tools for finding optimal alignment between two proteins, (3) cutting target protein into short sequence segments, (4) finding matched segments in databases according to the sequence alignment and the shape of template protein, (5) these segments co-ordinates are fitted to targeted structures until all atomic co-ordinates of the targeted structures are considered, (6) 10 times repeating these steps to generate average model and final energy minimisation for the entire targeted structure (Chou, 2004e).

Because of structural divergence of some high sequence similarity regions in proteins and alignment uncertainty in low complexity protein regions, even the best methods do not always generate highly accurate models. Introducing of “Consensus Server” (Prasad et al., 2004) can offer an improvement for comparative modeling by generating higher quality alignments.

SWISS-MODEL (Schwede et al., 2003) is one of the often used databases for homology modeling prediction.

10.1.1. SWISS-MODEL server

SWISS-MODEL is an automated comparative modeling server for prediction of 3D structure of proteins. The whole homology modelling steps can be handled with the ‘project mode’ using DeepView (Swiss-PdbViewer), an integrated sequence-structure workbench (Schwede et al., 2003). The results are being sent back via e-mail. This server is available under <http://swissmodel.expasy.org>.

By using the SWISS-MODEL server, we could get suitable 3D structure for five out of eight HPs. For HP Q9Y512, HP Q96I26 and HP Q9JG2 no reliable 3D template was predicted. In case of HP Q96I26, because of shortness of this protein no 3D structure could be expected. 1U0N (PDB) chain D was the best template for HP Q9BTR7 with 45% identity and an expectation value of $2e-08$. This template protein, is the ternary von Willebrand factor A1-glycoprotein Ib alpha-Botrocetin complex, characterised by X-ray diffraction and involved in blood clotting.

The solution structure of the RNA binding domain from mouse HP bab23670 (1W16; PDB; chain A) was a suitable

template for HP Q8IY67. This protein contains a RNA binding motif and belongs to nucleic acid binding proteins.

HP Q9POJ3 showed about 52% sequence identity to c1uc2a 3D structure template with 480 residues. This template, hypothetical extein protein of ph1602, was experimentally revealing endonuclease activity. Database results of HP Q9POJ3 showed that this protein may have similar structure and probably common function. From functional analysis searches (see above) we could not get any information regarding to function of this HP.

The 3D template for HP Q9BT99 is the C1 domain of Nore1, a novel Ras Effector (1rfh (PDB) chain A) (Fig. 7), which belongs to metal binding proteins.

10.2. Threading methods

Matching a sequence to a structure is so-called “threading”, that emerged in 1990s and is one of the methods for structural prediction of proteins without significant identity to already known structures (Hendlich et al., 1990; Miyazawa and Jernigan, 1996).

The threading method uses two different approximations (frozen and defrosted) to find similarity between evolutionary distant or even unrelated proteins (<30% sequence identity). In the frozen method the surrounding structural environments for each residue of the query are kept identical to those observed in the template structure. In contrast, the defrosted approach updates the surrounding amino acids of the template with the aligned amino acids of the query protein when calculating the fitness of the central residue (Bryant and Lawrence, 1993). Nevertheless, methods using the defrosted approach are much more accurate in predicting the fold of a protein than threading with frozen approximation. Fold recognition in CASP-3 (Moult et al., 1999) in 1998 was dominated by such methods (Panchenko et al., 1999).

10.3. Hybrid methods

Sequence similarity is not necessary for structural similarity, suggesting that convergent evolution can drive completely unrelated proteins to adopt the same fold. For better structural prediction the “hybrid-method” is used. This method combines sequence similarity with threading. The fast growth of the sequence databases contributed to a critical review of threading concept, which was based on the assumption that the local structural environment has an effect on the amino acid substitution pattern of each considered residue. Hybrid methods utilise sequence information from multiple sequence alignments if available, but also add characteristics like residue-based secondary structure preferences or preferences buried in the core of the protein. Because secondary structure or the pattern of exposed and buried residues in a structurally similar protein shows smaller variation than the amino acids and the local structural environments, these methods use “frozen approximation” for sequence alignment. Using dynamic programming matrix of scores and considering different contact potentials of the initial model of targeted protein

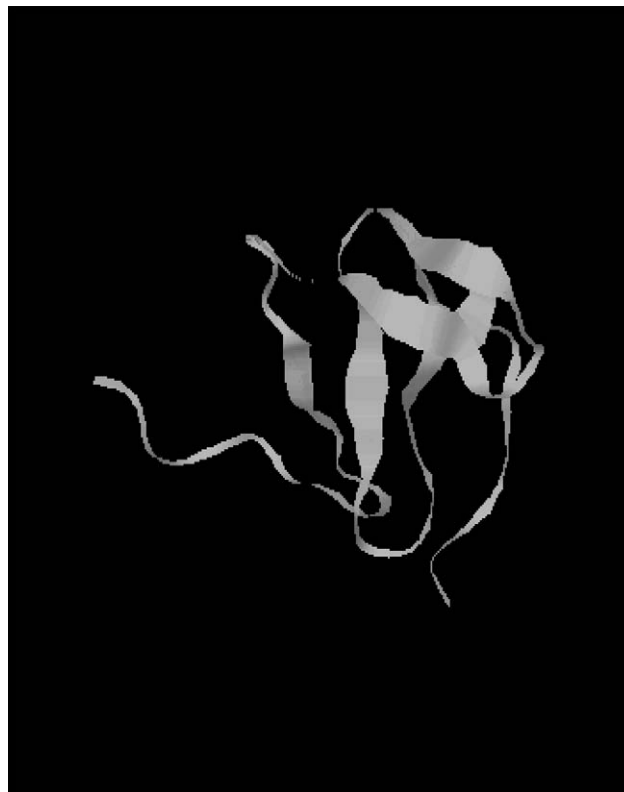


Fig. 7. Predicted 3D structure of HP Q9BT99 based on SWISS-MODEL server. The template protein is C1 domain of Nore1 (a novel Ras effector) with about 88% identity to target HP Q9BT99. The template length is 390 residue and expectation value of $3e-26$.

allow higher fold assignment rate than using information from routine PSI-BLAST. However, direct comparison methods like profile-profile alignment tools will be used favourably in protein structure prediction (Ginalski et al., 2005).

3D-PSSM (three-dimensional position-specific scoring matrix), is an example for a hybrid method. This method combines the power of multiple sequence profiles with knowledge of protein structure. It uses structural alignments of homologous proteins of similar three-dimensional structure in the structural classification of proteins (SCOP) database to get structural equivalence of residues. The resulting superfamily-based multiple alignment is conserved in a position specific matrix. 3D-PSSM allows structure-function relationships by combination sequence alignment methods with secondary structure matching and solvation potentials. This database is available at (<http://www.bmm.icnet.uk/servers/3dpssm>).

We used this server for prediction of 3D structure of HPs (Table 7) and these results were comparable to SWISS-MODEL-results except for HP Q9BTR7, HP Q8IY67 and HP Q9BT99: The 3D template protein of HP Q9BTR7 is an outer protein yopm protein which is a hydrolase/trypsin-like serine protease (Table 7). In the case of HP Q8IY67 the 3D template hud and Au-rich element of the tumor necrosis factor alphaRNA (1g2e: PDB) that was used, is involved in a transcription pathway (Table 7).

Table 7
3D-PSSM^a database search results of identified HPs from the medulloblastoma cell line

Accession number	Hypothetical protein name	Fold library	Template length	PSSM E_value	Fold	Superfamily	Template protein (PDB ID)
Q9BTR7	FLJ20331 protein [Fragment]	c1j15a (23% identity)	353	0.0388	PDB header: toxin	PDB: Chain: A: outer protein yopm	Novel molecular architecture of yopm-a leucine-rich2 effector protein from yersinia pestis (1j15)
Q8IY67	RAVER1	c1g2ea (25% identity)	167	0.000139	PDB header: transcription/rna	PDB Chain: A: Molecule:paraneoplastic encephalomyelitis antigen hud,.	Crystal structure of hud and au-rich element of the tumor2 necrosis factor alpha rna (1g2e)
Q96GS8	Hypothetical protein ABCF3	c1mt0a (24% identity)	241	0.0311	PDB header: Alpha and beta proteins (a/b)	P-loop containing nucleotide triphosphate hydrolases	P-loop containing nucleotide triphosphate hydrolases (1mt0)
Q9Y512	SAM50-like protein CGI-51	Not significant	–	–	–	–	–
Q96I26	PTPN11 protein	not significant	–	–	–	–	–
Q9P0J3	Putative 55 kDa protein	c1uc2 (50% identity)	480	6.42e–91	Structural genomics, unknown function.	Chain: A: PDB Molecule:hypothetical protein ph1602.	Hypothetical extein protein of ph1602 from pyrococcus2 horikoshii (1uc2)
Q9JJG2	Mus musculus brain cDNA, clone MNCb-2622, similar to AB033102 KIAA1276 protein	Not significant	–	–	–	–	–
Q9BT99	Ras association (RalGDS/AF-6) domain family 5, isoform D	d1tbo (21% identity)	66	7.58e–07	Protein kinase cystein-rich domain (cys2, phorbol-binding domain)	Protein kinase cystein-rich domain (cys2, phorbol-binding domain)	Protein kinase c-gamma (rat) family: Protein kinase cystein-rich domain (cys2, phorbol-binding domain) (1tbo)

^a Three-dimensional position specific matrix.

Finally, a protein kinase C-gamma phorbol-binding domain is a 3D template protein for HP Q9BT99 and reveals calcium binding function (Table 7).

10.4. Practical ab initio methods

According to various benchmarks, for 50% of cases fold recognition methods are not able to choose the correct fold for proteins without significant sequence similarity in databases. These methods have the limitation that no novel folds can be proposed.

The aims of ab initio methods are to find the native structure of a protein by simulating the biological process of protein folding. These methods perform iterative conformational changes and estimate the corresponding changes in energy. The inaccurate energy functions and vast number of possible chain conformations of a protein, are the main problems of this method. The reduced representation of conformations and inaccurate search strategies represent the second problem. By including lattice-based simulations of simplified protein models (Ortiz et al., 1999; Kolinski et al., 2003) and building structures from fragments of proteins (Simons et al., 1999; Bradley et al., 2003), more successful methods will be achieved. Another improvement of these methods is clustering of final conformations obtained after a large number of simulations. Representatives of large clusters are preferred as final models, which decrease the emphasis on calculated energy values.

Ab initio methods are increasingly applied in large-scale annotation projects, including fold assignments for small genomes and are also the only approach for designing new proteins (Dantas et al., 2003; Kuhlman et al., 2003). The scientist have to rely on pre-calculated results available, for example, for selected Pfam (Bateman et al., 2002) families. The ab initio methods are quite difficult to use and for interpretation of results expert knowledge is needed. Nevertheless, these methods are expected to play an important role in the future of structural biology.

11. Conclusion

In this communication we have indicated the importance of searching HPs. Fair generation of data using MS, following in-gel digestion of HPs with different proteolytic enzymes to increase sequence coverage, MS/MS and de novo sequencing for supporting the identification probability were shown and practical examples of methodologies were provided. Identification of HPs applying different databases was shown and recommendations for the use of databases were given or left open on purpose. The identified HPs were aligned in global and focal databases, homology searches were carried out; motif and domain searches in several programmes provided some clues for possible functions of HPs. Software options for subcellular localisation and signal peptide identifications were applied and complementing strategies as determination of physicochemical properties including amino acid composition and hydrophobicity scoring were addressed. We propose a prediction method

for membrane proteins and finally introduce some methods in conformational bioinformatics. Due to space limitation we were not in a position to individually discuss all methodologies and therefore this work only reflects an introduction into the matter, giving clues for determination and data mining. The review recommends a possible way to analyse this important and large group of proteins that occur in all proteomes but are frequently overlooked or ignored when proteomes are published. The paper as it stands is hermeneutical in nature and should therefore guide neurobiologists and neurochemists and should serve as an introduction by making essential and basic definitions. As many methodologies were not compared to others we do not claim that the recommended methods are representing the optimal technology, but in our hands methods given herein are forming the main practical approach to work on HPs. The practical examples listed were not completely discussed as they were meant to introduce the reader into the practical scenario of, thinking and strategies for HP proteomics. It was furthermore tried to introduce protein chemical know-how into neurobiology and neurochemistry and the reader is referred to a comprehensive reference list to allow going into details and further understanding. The eight demonstrated HPs have not been described elsewhere as we intended to show the use of methods on own original data, from MS, MS/MS, use of different proteases and de novo sequencing to extend sequence coverage and thus probability of identification. The tables are also shown to essentially demonstrate data mining, data handling and documentation according to protein chemical standards. Last not least, all predictions have to be experimentally verified and there is no information available in how many cases computational outcome was verified by experiments. We ask more advanced scientists for understanding when most of the contents is known to them and seems redundant.

Acknowledgements

We acknowledge supply of DAOY cells by Professor Dr. Irene Slavic (Medical University of Vienna, Dpt of Pediatrics) and Professor Dr. Thomas Stroebel (Medical University of Vienna, Clinical Institute for Neurology). We appreciate the kind assistance of Joo-Ho Shin, MSc for skilfull assistance with preparation of gels.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at [doi:10.1016/j.pneurobio.2005.10.001](https://doi.org/10.1016/j.pneurobio.2005.10.001).

References

- Afjehi-Sadat, L., Krapfenbauer, K., Slavic, I., Fountoulakis, M., Lubec, G., 2004. Hypothetical proteins with putative enzyme activity in human amnion, lymphocyte, bronchial epithelial and kidney cell lines. *Biochim. Biophys. Acta* 1700, 65–74.

- Afjeji-Sadat, L., Shin, J.H., Felizardo, M., Lee, K., Slavc, I., Lubec, G., 2005. Detection of hypothetical proteins in 10 individual human tumor cell lines. *Biochim. Biophys. Acta* 1747, 67–80.
- Ahram, M., Springer, D.L., 2004. Large-scale proteomic analysis of membrane proteins. *Expert Rev. Proteomics* 1, 293–302.
- Alfarano, C., Andrade, C.E., Anthony, K., Bahroos, N., Bajec, M., Bantoft, K., Betel, D., Bobechko, B., Boutilier, K., Burgess, E., Buzadzija, K., Cavero, R., D'Abreo, C., Donaldson, I., Dorairajoo, D., Dumontier, M.J., Dumontier, M.R., Earles, V., Farrall, R., Feldman, H., Garderman, E., Gong, Y., Gonzaga, R., Grytsan, V., Gryz, E., Gu, V., Haldorsen, E., Halupa, A., Haw, R., Hrvojic, A., Hurrell, L., Isserlin, R., Jack, F., Juma, F., Khan, A., Kon, T., Konopinsky, S., Le, V., Lee, E., Ling, S., Magidin, M., Moniakis, J., Montoj, J., Moore, S., Muskat, B., Ng, I., Paraiso, J.P., Parker, B., Pintilie, G., Pirone, R., Salama, J.J., Sgro, S., Shan, T., Shu, Y., Siew, J., Skinner, D., Snyder, K., Stasiuk, R., Strumpf, D., Tuekam, B., Tao, S., Wang, Z., White, M., Willis, R., Wolting, C., Wong, S., Wrong, A., Xin, C., Yao, R., Yates, B., Zhang, S., Zheng, K., Pawson, T., Ouellette, B.F., Hogue, C.W., 2005. The Biomolecular Interaction Network Database and related tools 2005 update. *Nucleic Acids Res.* 33, D418–D424.
- Altman, R.B., Dugan, J.M., 2003. Defining bioinformatics and structural bioinformatics. *Methods Biochem. Anal.* 44, 3–14.
- Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J., 1990. Basic local alignment search tool. *J. Mol. Biol.* 215, 403–410.
- Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J., 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25, 3389–3402.
- Altschul, S.F., Koonin, E.V., 1998. Iterated profile searches with PSI-BLAST—a tool for discovery in protein databases. *Trends Biochem. Sci.* 23, 444–447.
- Andreeva, A., Howorth, D., Brenner, S.E., Hubbard, T.J., Clothia, C., Murzin, A.G., 2004. SCOP database in 2004: refinements integrate structure and sequence family data. *Nucleic Acids Res.* 32, D226–D229.
- Apic, G., Huber, W., Teichmann, S.A., 2003. Multi-domain protein families and domain pairs: comparison with known structures and a random model of domain recombination. *J. Struct. Funct. Genomics* 4, 67–78.
- Attwood, T.K., Croning, M.D., Flower, D.R., Lewis, A.P., Mabey, J.E., Scordis, P., Selley, J.N., Wright, W., 2000. PRINTS-S: the database formerly known as PRINTS. *Nucleic Acids Res.* 28, 225–227.
- Attwood, T.K., 2002. The PRINTS database: a resource for identification of protein families. *Brief Bioinform.* 3, 252–263.
- Bairoch, A., Bucher, P., 1994. PROSITE: recent developments. *Nucleic Acids Res.* 22, 3583–3589.
- Bastolla, U., Porto, M., Roman, H.E., Vendruscolo, M., 2005. Looking at structure, stability, and evolution of proteins through the principal eigenvector of contact matrices and hydrophobicity profiles. *Gene* 347, 219–230.
- Bateman, A., Birney, E., Cerruti, L., Durbin, R., Ewinger, L., Eddy, S.R., Griffiths-Jones, S., Howe, K.L., Marshall, M., Sonnhammer, E.L., 2002. The Pfam protein families database. *Nucleic Acids Res.* 30, 276–280.
- Bendtsen, J.D., Nielsen, H., von Heijne, G., Brunak, S., 2004. Improved prediction of signal peptides: SignalP 3.0. *J. Mol. Biol.* 340, 783–795.
- Berndt, P., Hobohm, U., Langen, H., 1999. Reliable automatic protein identification from matrix-assisted laser desorption/ionization mass spectrometric peptide fingerprints. *Electrophoresis* 20, 3521–3526.
- Bimpikis, K., Budd, A., Linding, R., Gibson, T.J., 2003. BLAST2SRS, a web server for flexible retrieval of related protein sequences in the SWISS-PROT and SPTREMBL databases. *Nucleic Acids Res.* 31, 3792–3794.
- Bork, P., Ouzounis, C., Sander, C., Scharf, M., Schneider, R., Sonnhammer, E., 1992. What's in a genome? *Nature* 358, 287.
- Bradford, M.M., 1976. A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding. *Anal. Biochem.* 72, 248–254.
- Bradley, P., Chivian, D., Meiler, J., Misura, K.M., Rohl, C.A., Schief, W.R., Wedermeyer, W.J., Schueler-Furman, O., Murphy, P., Schonbrun, J., Strauss, C.E., Baker, D., 2003. Rosetta predictions in CASP5: successes, failures, and prospects for complete automation. *Proteins* 53, 457–468.
- Brenner, S.E., Chothia, C., Hubbard, T.J., Murzin, A.G., 1996. Understanding protein structure: using scop for fold interpretation. *Methods Enzymol.* 266, 635–643.
- Brenner, S.E., 2000. Target selection for structural genomics. *Nat. Struct. Biol.* 7 (Suppl), 967–969.
- Bru, C., Courcelle, E., Carrere, S., Beausse, Y., Dalmar, S., Kahn, D., 2005. The ProDom database of protein domain families: more emphasis on 3D. *Nucleic Acids Res.* 33, D212–D215.
- Bryant, S.H., Lawrence, C.E., 1993. An empirical energy function for threading protein sequence through the folding motif. *Proteins* 16, 92–112.
- Cedano, J., Aloy, P., Perez-Pons, J.A., Querol, E., 1997. Relation between amino acid composition and cellular location of proteins. *Mol. Biol.* 266, 594–600.
- Chamrad, D.C., Korting, G., Stuhler, K., Meyer, H.E., Klose, J., Bluggel, M., 2004. Evaluation of algorithms for protein identification from sequence databases using mass spectrometry data. *Proteomics* 4, 619–628.
- Chandonia, J.M., Walker, N.S., Lo Conte, L., Koehl, P., Levitt, M., Brenner, S.E., 2002. ASTRAL compendium enhancements. *Nucleic Acids Res.* 30, 260–263.
- Chothia, C., Lesk, A.M., 1986. The relation between the divergence of sequence and structure in proteins. *EMBO J.* 5, 823–826.
- Chothia, C., 1992. Proteins. One thousand families for the molecular biologist. *Nature* 357, 543–544.
- Chou, K.C., Jones, D., Heinrichson, R.L., 1997. Prediction of the tertiary structure and substrate binding site of caspase-8. *FEBS Lett.* 419, 49–54.
- Chou, J.J., Matsuo, H., Duan, H., Wagner, G., 1998. Solution structure of the RAID CARD and model for CARD/CARD interaction in caspase-2 and caspase-9 recruitment. *Cell* 94, 171–180.
- Chou, K.C., Watenpuugh, K.D., Heinrichson, R.L., 1999. A model of the complex between cyclin-dependent kinase 5 and the activation domain of neuronal Cdk5 activator. *Biochem. Biophys. Res. Commun.* 259, 420–428.
- Chou, K.C., 2000a. Prediction of protein structural classes and subcellular locations. *Curr. Protein Pept. Sci.* 1, 171–208.
- Chou, K.C., 2000b. Prediction of protein subcellular locations by incorporating quasi-sequence-order effect. *Biochem. Biophys. Res. Commun.* 278, 477–483.
- Chou, K.C., Tomasselli, A.G., Heinrichson, R.L., 2000. Prediction of the tertiary structure of a caspase-9/inhibitor complex. *FEBS Lett.* 470, 249–256.
- Chou, K.C., 2001a. Prediction of protein cellular attributes using pseudo-amino acid composition. *Proteins* 43, 246–255.
- Chou, K.C., 2001b. Prediction of signal peptides using scaled window. *Peptides* 22, 1973–1979.
- Chou, K.C., 2002. Prediction of protein signal sequences. *Curr. Protein Pept. Sci.* 3, 615–622.
- Chou, K.C., Cai, Y.D., 2002. Using functional domain composition and support vector machines for prediction of protein subcellular location. *J. Biol. Chem.* 277, 45765–45769.
- Chou, K.C., Howe, W.J., 2002. Prediction of the tertiary structure of the beta-secretase zymogen. *Biochem. Biophys. Res. Commun.* 292, 702–708.
- Chou, K.C., Cai, Y.D., 2003. A new hybrid approach to predict subcellular localization of proteins by incorporating gene ontology. *Biochem. Biophys. Res. Commun.* 311, 743–747.
- Chou, K.C., Wei, D.Q., Zhong, W.Z., 2003. Binding mechanism of coronavirus main proteinase with ligands and its implication to drug design against SARS. *Biochem. Biophys. Res. Commun.* 308, 148–151.
- Chou, K.C., 2004a. Insights from modelling the 3D structure of the extracellular domain of alpha7 nicotinic acetylcholine receptor. *Biochem. Biophys. Res. Commun.* 319, 433–438.
- Chou, K.C., 2004b. Insights from modeling the tertiary structure of human BACE2. *J. Proteome Res.* 3, 1069–1072.
- Chou, K.C., 2004c. Insights from modeling three-dimensional structures of the human potassium and sodium channels. *J. Proteome Res.* 3, 856–861.
- Chou, K.C., 2004d. Modelling extracellular domains of GABA-A receptors: subtypes 1, 2, 3, and 5. *Biochem. Biophys. Res. Commun.* 316, 636–642.
- Chou, K.C., 2004e. Structural bioinformatics and its impact to biomedical science. *Curr. Med. Chem.* 11, 2105–2134.
- Chou, K.C., 2005. Modeling the tertiary structure of human cathepsin-E. *Biochem. Biophys. Res. Commun.* 331, 56–60.
- Chou, K.C., Cai, Y.D., 2005a. Predicting protein localization in budding yeast. *Bioinformatics* 21, 944–950.

- Chou, K.C., Cai, Y.D., 2005b. Prediction of membrane protein types by incorporating amphipathic effects. *J. Chem. Inf. Model.* 45, 407–413.
- Chou, K.C., Cai, Y.D., 2005c. Using GO-PseAA predictor to identify membrane proteins and their types. *Biochem. Biophys. Res. Commun.* 327, 845–847.
- Chou, K.C., Elrod, D.W., 1999a. Protein subcellular location prediction. *Protein Eng.* 12, 107–118.
- Chou, K.C., Elrod, D.W., 1999b. Prediction of membrane protein types and subcellular locations. *Proteins* 34, 137–153.
- Claros, M.G., Brunak, S., von Heijne, G., 1997. Prediction of N-terminal protein sorting signals. *Curr. Opin. Struct. Biol.* 7, 394–398.
- Clauser, K.R., Baker, P., Burlingame, A.L., 1999. Role of accurate mass measurement (+/–10 ppm) in protein identification strategies employing MS or MS/MS and database searching. *Anal. Chem.* 71, 2871–2882.
- Coin, L., Bateman, A., Durbin, R., 2003. Enhanced protein domain discovery by using language modeling techniques from speech recognition. *Proc. Natl. Acad. Sci. USA* 100, 4516–4520.
- Dantas, G., Kuhlman, B., Callender, D., Wong, M., Baker, D., 2003. A large scale test of computational protein design: folding and stability of nine completely redesigned globular proteins. *J. Mol. Biol.* 332, 449–460.
- Doerks, T., Copley, R.R., Schultz, J., Ponting, C.P., Bork, P., 2002. Systematic identification of novel protein domain families associated with nuclear functions. *Genome Res.* 12, 47–56.
- Dokholyan, N.V., Shakhnovich, E.I., 2001. Understanding hierarchical protein evolution from first principles. *J. Mol. Biol.* 312, 289–307.
- Du, Q.S., Wang, S.Q., Zhu, Y., Wei, D.Q., Guo, H., Sirois, S., Chou, K.C., 2004. Polyprotein cleavage mechanism of SARS CoV Mpro and chemical modification of the octapeptide. *Peptides* 25, 1857–1864.
- Du, Q.S., Wang, S., Wei, D., Sirois, S., Chou, K.C., 2005. Molecular modeling and chemical modification for finding peptide inhibitor against severe acute respiratory syndrome coronavirus main proteinase. *Anal. Biochem.* 337, 262–270.
- Eriksson, J., Chait, B.T., Fenyo, D., 2000. A statistical basis for testing the significance of mass spectrometric protein identification results. *Anal. Chem.* 72, 999–1005.
- Falquet, L., Pagni, M., Bucher, P., Hulo, N., Sigrist, C.J., Hofmann, K., Bairoch, A., 2002. The PROSITE database, its status in 2002. *Nucleic Acids Res.* 30, 235–238.
- Fitch, W.M., 1970. Distinguishing homologous from analogous proteins. *Syst. Zool.* 19, 99–113.
- Galperin, M.Y., 2001. Conserved ‘hypothetical’ proteins: new hints and new puzzles. *Comp. Funct. Genomics* 2, 14–18.
- Galperin, M.Y., Koonin, E.V., 2004. ‘Conserved hypothetical’ proteins: prioritization of targets for experimental study. *Nucleic Acids Res.* 32, 5452–5463.
- Garg, A., Bhasin, M., Raghava, G.P., 2005. Support vector machine-based method for subcellular localization of human proteins using amino acid compositions, their order and similarity search. *J. Biol. Chem.* 280, 14427–14432.
- Ginalski, K., Grishin, N.V., Godzik, A., Rychlewski, L., 2005. Practical lessons from protein structure prediction. *Nucleic Acids Res.* 33, 1874–1891.
- Goldsmith-Fischman, S., Honig, B., 2003. Structural genomics: computational methods for structure analysis. *Protein Sci.* 12, 1813–1821.
- Gough, J., Karplus, K., Hughey, R., Chothia, C., 2001. Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure. *J. Mol. Biol.* 313, 903–919.
- Gras, R., Muller, M., Gasteiger, E., Gay, S., Binz, P.A., Bienvenut, W., Hoogland, C., Sanchez, J.C., Bairoch, A., Hochstrasser, D.F., Appel, R.D., 1999. Improving protein identification from peptide mass fingerprinting through a parameterized multi-level scoring algorithm and an optimized peak detection. *Electrophoresis* 20, 3535–3550.
- Guruprasad, K., Reddy, B.V., Pandit, M.W., 1990. Correlation between stability of a protein and its dipeptide composition: a novel approach for predicting in vivo stability of a protein from its primary sequence. *Protein Eng.* 4, 155–161.
- Hendlich, M., Lackner, P., Weitckus, S., Floeckner, H., Froschauer, R., Gottsbacher, K., Casari, G., Sippl, M.J., 1990. Identification of native protein folds amongst a large number of incorrect models. The calculation of low energy conformations from potentials of mean force. *J. Mol. Biol.* 216, 167–180.
- Henzel, W.J., Billeci, T.M., Stults, J.T., Wong, S.C., Grimley, C., Watanabe, C., 1993. Identifying proteins from two-dimensional gels by molecular mass searching of peptide fragments in protein sequence databases. *Proc. Natl. Acad. Sci. USA* 90, 5011–5015.
- Hillig, R.C., Renault, L., Vetter, I.R., Drell 4th, T., Wittinghofer, A., Becker, J., 1999. The crystal structure of rna1p: a new fold for a GTPase-activating protein. *Mol. Cell.* 3, 781–791.
- Holm, L., Sander, C., 1996. The FSSP database: fold classification based on structure-structure alignment of proteins. *Nucleic Acids Res.* 24, 206–209.
- Hua, S., Sun, Z., 2001. Support vector machine approach for protein subcellular localization prediction. *Bioinformatics* 17, 721–728.
- Hubbard, T.J., Ailey, B., Brenner, S.E., Murzin, A.G., Chothia, C., 1999. SCOP: a Structural Classification of Proteins database. *Nucleic Acids Res.* 27, 254–256.
- Ikai, A., 1980. Thermostability and aliphatic index of globular proteins. *J. Biochem.* 88, 1895–1898.
- Jacobsen, P.F., Jenkyn, D.J., Papadimitriou, J.M., 1985. Establishment of a human medulloblastoma cell line and its heterotransplantation into nude mice. *J. Neuropathol. Exp. Neurol.* 44, 472–485.
- James, P., Quadroni, M., Carafoli, E., Gonnet, G., 1993. Protein identification by mass profile fingerprinting. *Biochem. Biophys. Res. Commun.* 195, 58–64.
- Junker, V.L., Apweiler, R., Bairoch, A., 1999. Representation of functional information in the SWISS-PROT data bank. *Bioinformatics* 15, 1066–1067.
- Kauzmann, W., 1959. Some factors in the interpretation of protein denaturation. *Adv. Protein Chem.* 14, 1–63.
- Kolesov, G., Mewes, H.W., Frishman, D., 2002. SNAPper: gene order predicts gene function. *Bioinformatics* 18, 1017–1019.
- Kolinski, A., Gront, D., Pokarowski, P., Skolnick, J., 2003. A simple lattice model that exhibits a rotein-like cooperative all-or-none folding transition. *Biopolymers* 69, 399–405.
- Koonin, E.V., Galperin, M.Y., 2003. Protein sequence motifs and domain databases. In: Koonin, E.V., Galperin, M.Y. (Eds.), *Sequence-evolution-function: Computational Approaches in Comparative Genomics*. Kluwer Academic Publishers.
- Krogh, A., Larsson, B., von Heijne, G., Sonnhammer, E.L., 2001. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol.* 305, 567–580.
- Kuhlman, B., Dantas, G., Ireton, G.C., Varani, G., Stoddard, B.L., Baker, D., 2003. Design of a novel globular protein fold with atomic-level accuracy. *Science* 302, 1364–1368.
- Kyte, J., Doolittle, R.F., 1982. A simple method for displaying the hydrophobic character of a protein. *J. Mol. Biol.* 157, 105–132.
- Lo Conte, L., Brenner, S.E., Hubbard, T.J., Chothia, C., Murzin, A.G., 2002. SCOP database in 2002: refinements accommodate structural genomics. *Nucleic Acids Res.* 30, 264–267.
- Lesk, A.M., 2003. Hydrophobicity-getting into hot water. *Biophys. Chem.* 105, 179–182.
- Letunic, I., Copley, R.R., Schmidt, S., Ciccarelli, F.D., Doerks, T., Schultz, J., Ponting, C.P., Bork, P., 2004. SMART 4.0: towards genomic data integration. *Nucleic Acids Res.* 32, D142–D144.
- Lipman, D.J., Pearson, W.R., 1985. Rapid and sensitive protein similarity searches. *Science* 227, 1435–1441.
- Lubec, G., Krapfenbauer, K., Fountoulakis, M., 2003. Proteomics in brain research: potentials and limitations. *Prog. Neurobiol.* 69, 193–211.
- Marchler-Bauer, A., Bryant, S.H., 1997. A measure of success in fold recognition. *Trends Biochem. Sci.* 22, 236–240.
- Marchler-Bauer, A., Bryant, S.H., 2004. CD-Search: protein domain annotations on the fly. *Nucleic Acids Res.* 32, W327–W331.
- Marcus, K., Immmler, D., Sternberger, J., Meyer, H.E., 2000. Identification of platelet proteins separated by two-dimensional gel electrophoresis and analyzed by matrix assisted laser desorption/ionization-time of flight-mass spectrometry and detection of tyrosine-phosphorylated proteins. *Electrophoresis* 21, 2622–2636.

- Mellor, J.C., Yanai, I., Clodfelter, K.H., Mintseris, J., DeLisi, C., 2002. Predictome: a database of putative functional links between proteins. *Nucleic Acids Res.* 30, 306–309.
- Minion, F.C., Lefkowitz, E.J., Madsen, M.L., Cleary, B.J., Swartzell, S.M., Mahairas, G.G., 2004. The genome sequence of *Mycoplasma hyopneumoniae* strain 232, the agent of swine mycoplasmosis. *J. Bacteriol.* 186, 7123–7133.
- Miyazawa, S., Jernigan, R.L., 1996. Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. *J. Mol. Biol.* 256, 623–644.
- Mizuguchi, K., Deane, C.M., Blundell, T.L., Overington, J.P., 1998. HOMSTRAD: a database of protein structure alignments for homologous families. *Protein Sci.* 7, 2469–2471.
- Mosimann, S., Meleshko, R., James, M.N., 1995. A critical assessment of comparative molecular modeling of tertiary structures of proteins. *Proteins* 23, 301–317.
- Mulder, N.J., Apweiler, R., Attwood, T.K., Bairoch, A., Bateman, A., Binns, D., Bradley, P., Bork, P., Bucher, P., Cerutti, L., Copley, R., Courcelle, E., Das, U., Durbin, R., Fleischmann, W., Gough, J., Haft, D., Harte, N., Hulo, N., Kahn, D., Kanapin, A., Krestyaninova, M., Lonsdale, D., Lopez, R., Letunic, I., Madera, M., Maslen, J., McDowall, J., Mitchell, A., Nikolskaya, A.N., Orchard, S., Pagni, M., Ponting, C.P., Quevillon, E., Selengut, J., Sigrist, C.J., Silventoinen, V., Studholme, D.J., Vaughan, R., Wu, C.H., 2005. InterPro, progress and status in 2005. *Nucleic Acids Res.* 33, D201–D205.
- Murzin, A.G., Brenner, S.E., Hubbard, T., Clothia, C., 1995. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* 247, 536–540.
- Nakai, K., Kanehisa, M., 1992. A knowledge base for predicting protein localization sites in eukaryotic cells. *Genomics* 14, 897–911.
- Nakai, K., Horton, P., 1999. PSORT: a program for detecting the sorting signals of proteins and predicting their subcellular localization. *Trends Biochem. Sci.* 24, 34–36.
- Nakashima, H., Nishikawa, K., 1992. The amino acid composition is different between the cytoplasmic and extracellular sides in membrane proteins. *FEBS Lett.* 303, 141–146.
- Nakashima, H., Nishikawa, K., 1994. Discrimination of intracellular and extracellular proteins using amino acid composition and residue-pair frequencies. *J. Mol. Biol.* 238, 54–61.
- Nakashima, H., Nishikawa, K., Ooi, T., 1986. The folding type of a protein is relevant to the amino acid composition. *J. Biochem. (Tokyo)* 99, 153–162.
- Nandi, T., Kannan, K., Ramachandran, S., 2003. The low complexity proteins from enteric pathogenic bacteria: taxonomic parallels embedded in diversity. *In Silico Biol.* 3, 277–285.
- Needleman, S.B., Wunsch, C.D., 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* 48, 443–453.
- Nielsen, H., Engelbrecht, J., Brunak, S., von Heijne, G., 1997. Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Protein Eng.* 10, 1–6.
- Nishikawa, K., Kubota, Y., Ooi, T., 1983. Classification of proteins into groups based on amino acid composition and other characters. II. Grouping into four types. *J. Biochem. (Tokyo)* 94, 997–1007.
- Oh, J.E., Krapfenbauer, K., Fountoulakis, M., Frischer, T., Lubec, G., 2004. Evidence for the existence of hypothetical proteins in human bronchial epithelial, fibroblast, amnion, lymphocyte, mesothelial and kidney cell lines. *Amino Acids* 26, 9–18.
- Orengo, C.A., Pearl, F.M., Bray, J.E., Todd, A.E., Martin, A.C., Lo Conte, L., Thornton, J.M., 1999. The CATH Database provides insights into protein structure/function relationships. *Nucleic Acids Res.* 27, 275–279.
- Ortiz, A.R., Kolinski, A., Rotkiewicz, P., Ilkowski, B., Skolnick, J., 1999. Ab initio folding of proteins using restraints derived from evolutionary information. *Proteins Suppl.* 3, 177–185.
- Osguthorpe, D.J., 2000. Ab initio protein folding. *Curr. Opin. Struct. Biol.* 10, 146–152.
- Pan, Y.X., Zhang, Z.Z., Guo, Z.M., Feng, G.Y., Hunag, Z.D., He, L., 2003. Application of pseudo amino acid composition for predicting protein subcellular location: stochastic signal processing approach. *J. Protein Chem.* 22, 395–402.
- Panchenko, A., Marchler-Bauer, A., Bryant, S.H., 1999. Threading with explicit models for evolutionary conservation of structure and sequence. *Proteins (Suppl. 3)*, 133–140.
- Pandit, S.B., Bhadra, R., Gowri, V.S., Balaji, S., Anand, B., Srinivasan, N., 2004. SUPFAM: a database of sequence superfamilies of protein domains. *BMC Bioinform.* 5, 28.
- Pearl, F., Todd, A., Sillitoe, I., Dibley, M., Redfern, O., Lewis, T., Bennett, C., Marsden, R., Grant, A., Lee, D., Akpor, A., Maibaum, M., Harrison, A., Dallman, T., Reeves, G., Diboun, I., Addou, S., Lise, S., Johnston, C., Sillero, A., Thornton, J., Orengo, C., 2005. The CATH Domain Structure Database and related resources Gene3D and DHS provide comprehensive domain family information for genome analysis. *Nucleic Acids Res.* 33, D247–D251.
- Pearson, W.R., Lipman, D.J., 1988. Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci. USA* 85, 2444–2448.
- Perkins, D.N., Pappin, D.J., Creasy, D.M., Cottrell, J.S., 1999. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* 20, 3551–3567.
- Prasad, J.C., Vajda, S., Camacho, C.J., 2004. Consensus alignment server for reliable comparative modeling with distant templates. *Nucleic Acids Res.* 32, W50–W54.
- Puntervoll, P., Linding, R., Gemund, C., Chabanis-Davidson, S., Mattingsdal, M., Cameron, S., Martin, D.M., Ausiello, G., Brannetti, B., Costantini, A., Ferre, F., Maselli, V., Via, A., Cesareni, G., Diella, F., Superti-Furga, G., Wyrwicz, L., Ramu, C., McGuigan, C., Gudavalli, R., Letunic, I., Bork, P., Rychlewski, L., Kuster, B., Helmer-Citterich, M., Hunter, W.N., Aasland, R., Gibson, T.J., 2003. ELM server: A new resource for investigating short functional sites in modular eukaryotic proteins. *Nucleic Acids Res.* 31, 3625–3630.
- Reddy, B.V., 1996. Structural distribution of dipeptides that are identified to be determinants of intracellular protein stability. *J. Biomol. Struct. Dyn.* 14, 201–210.
- Reece, G.R., de Haen, C., Teller, D.C., Doolittle, R.F., Fitch, W.M., Dickerson, R.E., Chambon, P., McLachlan, A.D., Margoliash, E., Jukes, T.H., et al., 1987. Homology in proteins and nucleic acids: a terminology muddle and a way out of it. *Cell* 50, 667.
- Reinhardt, A., Hubbard, T., 1998. Using neural networks for prediction of the subcellular location of proteins. *Nucleic Acids Res.* 26, 2230–2236.
- Rost, B., 1996. PHD: predicting one-dimensional protein structure by profile-based neural networks. *Methods Enzymol.* 266, 525–539.
- Rost, B., 1999. Twilight zone of protein sequence alignments. *Protein Eng.* 12, 85–94.
- Rost, B., Valencia, A., 1996. Pitfalls of protein sequence analysis. *Curr. Opin. Biotechnol.* 7, 457–461.
- Sanchez, R., Sali, A., 1997. Advances in comparative protein-structure modeling. *Curr. Opin. Struct. Biol.* 7, 206–214.
- Saraf, M.C., Horswill, A.R., Benkovic, S.J., Maranas, C.D., 2004. FamClash: a method for ranking the activity of engineered enzymes. *Proc. Natl. Acad. Sci. USA* 101, 4142–4147.
- Sauder, J.M., Arthur, J.W., Dunbrack Jr., R.L., 2000. Large-scale comparison of protein sequence alignment algorithms with structure alignments. *Proteins* 40, 6–22.
- Schwede, T., Kopp, J., Guex, N., Peitsch, M.C., 2003. SWISS-MODEL: An automated protein homology-modeling server. *Nucleic Acids Res.* 31, 3381–3385.
- Shapiro, L., Harris, T., 2000. Finding function through structural genomics. *Curr. Opin. Biotechnol.* 11, 31–35.
- Shevchenko, A., Sunyaev, S., Loboda, A., Shevchenko, A., Bork, P., Ens, W., Standing, K.G., 2001. Charting the proteomes of organisms with unsequenced genomes by MALDI-quadrupole time-of-flight mass spectrometry and BLAST homology searching. *Anal. Chem.* 73, 1917–1926.
- Shin, J.H., Yang, J.W., Juranville, J.F., Fountoulakis, M., Lubec, G., 2004a. Evidence for existence of thirty hypothetical proteins in rat brain. *Proteome Sci.* 2, 1.

- Shin, J.H., Gulesserian, T., Weitzdoerfer, R., Fountoulakis, M., Lubec, G., 2004b. Derangement of hypothetical proteins in fetal Down's syndrome brain. *Neurochem. Res.* 29, 1307–1316.
- Shin, J.H., Yang, J.W., Le Pecheur, M., London, J., Hoeger, H., Lubec, G., 2004c. Altered expression of hypothetical proteins in hippocampus of transgenic mice overexpressing human Cu/Zn-superoxide dismutase 1. *Proteome Sci.* 2, 2.
- Siew, N., Fischer, D., 2003a. Analysis of singleton ORFans in fully sequenced microbial genomes. *Proteins* 53, 241–251.
- Siew, N., Fischer, D., 2003b. Twenty thousand ORFan microbial protein families for the biologist? *Structure* 11, 7–9.
- Simons, K.T., Bonneau, R., Ruczinski, I., Baker, D., 1999. Ab initio protein structure prediction of CASP III targets using ROSETTA. *Proteins (Suppl. 3)*, 171–176.
- Sirois, S., Wei, D.Q., Du, Q., Chou, K.C., 2004. Virtual screening for SARS-CoV protease based on KZ7088 pharmacophore points. *J. Chem. Inf. Comput. Sci.* 44, 1111–1122.
- Smith, T.F., Waterman, M.S., 1981. Identification of common molecular subsequences. *J. Mol. Biol.* 147, 195–197.
- Shmueli, H., Dinitz, E., Dahan, I., Eichler, J., Fischer, D., Shaanan, B., 2004. Poorly conserved ORFs in the genome of the archaea *Halobacterium sp* NRC-1 correspond to expressed proteins. *Bioinformatics* 20, 1248–1253.
- Sowdhamini, R., Burke, D.F., Huang, J.F., Mizuguchi, K., Nagarajaram, H.A., Srinivasan, N., Steward, R.E., Blundell, T.L., 1998. CAMPASS: a database of structurally aligned protein superfamilies. *Structure* 6, 1087–1094.
- Stevens, R., Goble, C.A., Bechhofer, S., 2000. Ontology-based knowledge representation for bioinformatics. *Brief Bioinform.* 1, 398–414.
- Stevens, F.J., 2005. Efficient recognition of protein fold at low sequence identity by conservative application of Psi-BLAST: validation. *J. Mol. Recognit.* 18, 139–149.
- Suckau, D., Resemann, A., Schuerenberg, M., Hufnagel, P., Franzen, J., Holle, A., 2003. A novel MALDI LIFT-TOF/TOF mass spectrometer for proteomics. *Anal. Bioanal. Chem.* 376, 952–965.
- Tang, C., Zhang, W., Fenyó, D., Chait, B.T., 2000. Assessing the Performance of Different Protein Identification Algorithms. In: 48TH ASMS Conference, June 11–15, Long Beach, California.
- Tatusov, R.L., Natale, D.A., Garkavtsev, I.V., Tatusova, T.A., Shankavaram, U.T., Rao, B.S., Kiryutin, B., Galperin, M.Y., Fedorova, N.D., Koonin, E.V., 2001. The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Res.* 29, 22–28.
- Tusnady, G.E., Simon, I., 2001. The HMMTOP transmembrane topology prediction server. *Bioinformatics* 17, 849–850.
- Veretnik, S., Bourne, P.E., Alexandrov, N.N., Shindyalov, I.N., 2004. Toward consistent assignment of structural domains in proteins. *J. Mol. Biol.* 339, 647–678.
- von Heijne, G., 1984. Analysis of the distribution of charged residues in the N-terminal region of signal sequences: implications for protein export in prokaryotic and eukaryotic cells. *EMBO J.* 3, 2315–2318.
- von Heijne, G., 1992. Membrane protein structure prediction. Hydrophobicity analysis and the positive-inside rule. *J. Mol. Biol.* 225, 487–494.
- von Mering, C., Huynen, M., Jaeggi, D., Schmidt, S., Bork, P., Snel, B., 2003. STRING: a database of predicted functional associations between proteins. *Nucleic Acids Res.* 31, 258–261.
- von Mering, C., Jensen, L.J., Snel, B., Hooper, S.D., Krupp, M., Foglierini, M., Jouffre, N., Huynen, M.A., Bork, P., 2005. STRING: known and predicted protein–protein associations, integrated and transferred across organisms. *Nucleic Acids Res.* 33, D433–D437.
- Wang, M., Yang, J., Liu, G.P., Xu, Z.J., Chou, K.C., 2004. Weighted-support vector machines for predicting membrane protein types based on pseudo-amino acid composition. *Protein Eng Des Sel.* 17, 509–516.
- Wang, M., Yang, J., Chou, K.C., 2005a. Using string kernel to predict signal peptide cleavage site based on subsite coupling model. *Amino Acids* 28, 395–402.
- Wang, M., Yang, J., Xu, Z.J., Chou, K.C., 2005b. SLLE for predicting membrane protein types. *J Theor Biol.* 232, 7–15.
- Weitzdoerfer, R., Fountoulakis, M., Lubec, G., 2002. Reduction of actin-related protein complex 2/3 in fetal Down syndrome brain. *Biochem. Biophys. Res. Commun.* 293, 836–841.
- Westbrook, J., Feng, Z., Jain, S., Bhat, T.N., Thanki, N., Ravichandran, V., Gilliland, G.L., Bluhm, W., Weissig, H., Greer, D.S., Bourne, P.E., Berman, H.M., 2002. The Protein Data Bank: unifying the archive. *Nucleic Acids Res.* 30, 245–248.
- Westbrook, J., Feng, Z., Chen, L., Yang, H., Berman, H.M., 2003. The Protein Data Bank and structural genomics. *Nucleic Acids Res.* 31, 489–491.
- Wolf, Y.I., Grishin, N.V., Koonin, E.V., 2000. Estimating the number of protein folds and families from complete genome data. *J. Mol. Biol.* 299, 897–905.
- Xenarios, I., Salwinski, L., Duan, X.J., Higney, P., Kim, S.M., Eisenberg, D., 2002. DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. *Nucleic Acids Res.* 30, 303–305.
- Xiao, X., Shao, S., Ding, Y., Huang, Y., Chou, K.C., 2005. Using complexity measure factor to predict protein subcellular location. *Amino Acids.* 28, 57–61.
- Yang, J.W., Czech, T., Lubec, G., 2004. Proteomic profiling of human hippocampus. *Electrophoresis* 25, 1169–1174.
- Yates 3rd, J.R., Speicher, S., Griffin, P.R., Hunkapiller, T., 1993. Peptide mass maps: a highly informative approach to protein identification. *Anal. Biochem.* 214, 397–408.
- Zhang, W., Chait, B.T., 2000. ProFound: an expert system for protein identification using mass spectrometric peptide mapping information. *Anal. Chem.* 72, 2482–2489.
- Zhou, G.P., Doctor, K., 2003. Subcellular location prediction of apoptosis proteins. *Proteins* 50, 44–48.
- Zhang, Z., Schaffer, A.A., Miller, W., Madden, T.L., Lipman, D.J., Koonin, E.V., Altschul, S.F., 1998. Protein sequence similarity searches using patterns as seeds. *Nucleic Acids Res.* 26, 3986–3990.
- Zhang, Z., Ng, S.K., 2004. InterWeaver: interaction reports for discovering potential protein interaction partners with online evidence. *Nucleic Acids Res.* 32, W73–W75.