BMC Genomics

Database

PCAS – a precomputed proteome annotation database resource

Yong Zhang, Yanbin Yin, Yunjia Chen, Ge Gao, Peng Yu, Jingchu Luo and Ying Jiang*

Address: College of Life Sciences, National Laboratory of Genetic Engineering and Protein Engineering, Center of Bioinformatics, Peking University, Beijing 100871, China

Email: Yong Zhang - zhangy@mail.cbi.pku.edu.cn; Yanbin Yin - yinyb@mail.cbi.pku.edu.cn; Yunjia Chen - chenyj@mail.cbi.pku.edu.cn;
Ge Gao - gaog@mail.cbi.pku.edu.cn; Peng Yu - yup@mail.cbi.pku.edu.cn; Jingchu Luo - luojc@mail.cbi.pku.edu.cn;
Ying Jiang* - ying.jiang@spcorp.com

* Corresponding author

Published: 01 November 2003

BMC Genomics 2003, **4**:42

Received: 12 August 2003
Accepted: 01 November 2003

This article is available from: http://www.biomedcentral.com/1471-2164/4/42

Abstract

Background: Many model proteomes or "complete" sets of proteins of given organisms are now publicly available. Much effort has been invested in computational annotation of those "draft" proteomes. Motif or domain based algorithms play a pivotal role in functional classification of proteins. Employing most available computational algorithms, mainly motif or domain recognition algorithms, we set up to develop an online proteome annotation system with integrated proteome annotation data to complement existing resources.

Results: We report here the development of PCAS (ProteinCentric Annotation System) as an online resource of pre-computed proteome annotation data. We applied most available motif or domain databases and their analysis methods, including *hmmpfam* search of HMMs in Pfam, SMART and TIGRFAM, *RPS-PSIBLAST* search of PSSMs in CDD, *pfscan* of PROSITE patterns and profiles, as well as *PSI-BLAST* search of SUPERFAMILY PSSMs. In addition, signal peptide and TM are predicted using *SignalP* and *TMHMM* respectively. We mapped SUPERFAMILY and COGs to *InterPro*, so the motif or domain databases are integrated through *InterPro*. PCAS displays table summaries of pre-computed data and a graphical presentation of motifs or domains relative to the protein. As of now, PCAS contains human IPI, mouse IPI, and rat IPI, *A. thaliana*, *C. elegans*, *D. melanogaster*, *S. cerevisiae*, and *S. pombe* proteome.

PCAS is available at http://pak.cbi.pku.edu.cn/proteome/gca.php

Conclusion: PCAS gives better annotation coverage for model proteomes by employing a wider collection of available algorithms. Besides presenting the most confident annotation data, PCAS also allows customized query so users can inspect statistically less significant boundary information as well. Therefore, besides providing general annotation information, PCAS could be used as a discovery platform. We plan to update PCAS twice a year. We will upgrade PCAS when new proteome annotation algorithms identified.

Background

Proteome is defined as a "complete" set of proteins of a given model organism. Many proteomes are now available. Much effort have been invested in computational annotation of those "draft" proteomes. Motif or domain based algorithms play pivotal roles in proteome annotation. This is mostly because of their sensitivity, coverage and the well-curated functional information associated with known motifs or domains. Over the years, many motif or domain databases have been developed (PROSITE [1], Pfam [2], SMART [3], TIGRFAM [4], PRINTS [5], Blocks [6], ProDom [7] etc.), so have their underlying computation methods (pfscan [8], HMMER [9], FingerPRINTScan [10] etc.). *InterPro* [11] provides an integrated resource to cross-reference such motif or domain databases. *InterProScan* [12] has integrated various analysis methods to provide an *InterPro* based integrated platform for motif or domain recognition. Consistent with this effort, CDD [13] provides another rich resource for functional motifs or domains, which could complement those in *InterPro*. RPS-PSIBLAST [14] is the analysis method to classify protein with CDD PSSMs. There are also regular expression pattern matching methods for general or special applications in database search [15-18]. However, they are less used in general proteome annotation effort.

Since different motif or domain databases and their specific algorithms, diagnostically may have different sensitivities and specificities, they usually have different coverages when annotating a collection of proteins such as proteome [19]. Thus, as *InterPro* [11,12] and PIR [20], integrated annotation platform best consists of a collection of motif or domain databases and their respective analysis methods. Complementary to existing resources, here we developed an online proteome annotation system with wider collection of available algorithms and coverage of integrated domain or motif databases.

Construction and Content

Materials and Methods

- **Proteomes:** SWISS-PROT/TrEMBL non-redundant complete proteome sets of *A. thaliana*, *C. elegans*, *D. melanogaster*, *S. cerevisiae*, and *S. pombe* were downloaded from EBI on Apr 25th, 2003; human IPI 2.18, mouse IPI 1.11, rat IPI 1.1 from EBI. Zebrafish from NCBIrefseq. Fugu and *Ciona inestinalis* from JGI

- **Domain or Motif Databases/Algorithms:** *InterPro* database version 6.2 is downloaded from EBI; CDD database version 1.62 is from NCBI; Blast Package 2.2.6 downloaded from NCBI; Superfamily 1.61 PSSMs downloaded from MRC-LMB; Printscan 3.595 and PRINTS 35.0 downloaded from UMBER; Prosite database 17.26 and search tool pfscan 1.5 downloaded from Expasy; Profile HMMER 2.3 downloaded from Wustl; SignalP V2.0.b2 from CBS; TMHMM 2.0 was run using CBS web service;

- **Software and Programming Languages:** Perl language 5.8.0 is from http://Perl.com; PHP 4.3.2 from http://php.net; MySQL 4.0.12 from http://mysql.com; Apache 1.3.26 from http://Apache.org; GD graphics library 1.8.0 from http://Boutell.com

All of the algorithms were run with default parameters. After that, the original results were parsed by Perl scripts. Data were stored in MySQL database.

Apache and PHP were used to set up the web server and write the web pages. The graphical display in PCAS was realized by calling GD graphics library.

Annotation Pipeline

PCAS, ProteinCentric Annotation System, includes most of the motif or domain databases and analysis methods mentioned above. Moreover, we included SCOP [21] based SUPERFAMILY [22] classification and COG [23] classification. SUPERFAMILY analysis offers structure based protein annotation; COG, short for Clusters of Orthologous Groups of proteins, is a system delineated by comparing protein sequences encoded in over forty complete genomes that represent 30 major phylogenetic lineages. Assigning proteins to COGs could provide not only valuable evolutional inference but also functional information derived from evolutional analysis. We also included SignalP [24] and TMHMM [25] to perform a priori prediction of leader peptide and TM regions, which is indicative of proper function of a protein.

Following are algorithms, motif or domain databases and application software in current PCAS:

hmmpfam V2.2g search of Pfam 8.0 , SMART 3.4, TIGERFAM 2.1 HMMs

RPS-BLAST V2.2.6 of CDD 1.62 PSSMs: COG, Pfam, Smart, LOAD and NCBI curated CD

pfscan V1.5 of PROSITE 17.26 patterns and profiles.

PSI-BLAST V2.2.6 of SCOP based SUPERFAMILY 1.61 PSSMs

FingerPRINTScan V3.595 of PRINTS 35.0 fingerprints.

SignalP V2.0.b2 for signal peptide prediction.

TMHMM 2.0 for TM prediction.

Cross-reference Motif or Domain Database

We used *InterPro* system as the basis for data integration, which consists of most available motif or domain databases. Certain subsets in CDD and *InterPro* are overlapping. Therefore, through CDD, we can also map COG and SUPERFAMILY to *InterPro*.

Mapping COGs to InterPro

According to the embedded relations between different CDD subsets, we can map COGs to Pfam and Smart then to *InterPro*. For example, in CDD database, COG0004 is related to Pfam00909, which is a signature of IPR001905; COG0004 is then mapped to IPR001905. Among the 4873 COG PSSMs in current CDD release 1.62, 2285 COGs were related to Pfam or SMART, and were mapped to *InterPro* database. Some of the unmapped COGs may be associated with more than one *InterPro* entry, and the rest simply are not included in current *InterPro* 6.1 yet.

Mapping SUPERFAMILY to InterPro

There are 7550 HMMs representing 1109 superfamilies. We first map those HMMs to CDDs by performing hmmpfam search against CDD's representative sequences, then using the embedded CDD and *InterPro* cross references, map those SUPERFAMILY HMMs to *InterPro* entries and further map SUPERFAMILY to *InterPro*. Detailed mapping strategy is described as following:

Step1, apply E value cutoff at 0.01.

202 SUPERFAMILY HMMs that have no hits and 464 HMMs that only have hits with E value greater than 0.01 were thrown out.

Step2: filter out those with CDD and *InterPro* link broken

90 HMMs are filtered out since their CDD hits have no corresponding *InterPro* entries.

The remaining 6994 (7550-202-464-90) HMMs were divided into 3548 1:1 relations (one SUPERFAMILY model corresponds to one *InterPro* entry) and 3446 1:N relations (one SUPERFAMILY model corresponds to more than one *InterPro* entries).

Step3: apply coverage filters

For 1:1 relations, we applied following coverage filters:

0.5 <= sf_length / cdd_length <= 2;

(The length ratio of SUPERFAMILY HMM and CDD representative sequence)

sf_coverage >= 0.75;

(The length coverage of the aligned query SUPERFAMILY HMM over the full length of the query SUPERFAMILY HMM.)

cdd_coverage >= 0.75.

(The length coverage of the aligned CDD representative sequence over the full length of the CDD representative sequence)

Total 561 out of 3446 1:1 relations were initially filtered out. By matching descriptions in SUPERFAMILY and *InterPro*, we further retained 4 mappings.

For 1:N relations, the coverage filters are a bit more stringent that above:

0.8 <=sf_length/cdd_length < = 1.25

sf_coverage > = 0.75

cdd_coverage > = 0.75

1957 1:N relations were filtered out. By matching descriptions in SUPERFAMILY and *InterPro*, we further retained 22 mappings. At this point, all the mappings are 1:1 relations.

In summary, 4502 (26+1489+2987) out of 7550 SUPERFAMILY HMMs are mapped to *InterPro* entries, and they are all 1:1 mappings. In terms of 1109 superfamilies, 551 have only one *InterPro* entry; the rest are mapped to zero or multiple *InterPro* entries, with 66 as the maximum (P-loop containing nucleotide triphosphate hydrolases in SUPERFAMILY). For those with multiple *InterPro* entries, we treated them as unmapped, since those *InterPro* entries most likely represent certain subfamilies and it is scientifically inaccurate to use subfamily's definition to describe superfamily.

We would like to point out that the purpose of mapping SUPERFAMILY to *InterPro* here is to find out the best possible *InterPro* description for a particular SUPERFAMILY member. Therefore, we took relatively stringent mapping conditions described above. This enables PCAS to steer clear from the possible complication caused by multi-domain protein families or by superfamily, subfamily relationships.

Utility and Discussion

PCAS Query and Display

The query of PCAS for pre-computed proteome annotation data is initiated with a protein identifier, for example, the SPTR or IPI protein ID or AC. One can also query PCAS through key word text search or blast with protein amino acid sequence. When blast is used, lists of blast hits in the target proteomes will be returned if exist, which are linked to the pre-computed annotation information. One needs to infer the annotation of the query protein from the blast hits. However, this annotation transfer should not happen if the query protein is not similar enough to the blast hits [26,27]. The query is customized so users can specify the statistical cut off (not shown); users can even directly specify the number of hits to be displayed (5 is the default). Query customization in PCAS allows query for statistically less confident borderline annotation data which may be indicative of novel functions.

Figure 1 shows the PCAS query display, which is organized as following. General header is the header information of the protein in the proteome. It comes from the public annotation effort when the proteome is released.

After the header information, there is the overview of the computational results, which consists of table summaries of pre-computed annotation data. The first table summary is the pre-computed data from motif or domain based algorithms and sorted based on statistical significance. In this table format, *InterPro* based data integration is expressed by coloring. Hits mapped to the same IPR name (*InterPro* ID) will be displayed in the same color. In another word, the same motif or domain hit by the query protein using different algorithms is displayed with the same color. Currently the color-coding is limited to the top five different protein motifs or domains for simplicity. This color-coding schema is implemented throughout this display page including the *InterPro* description table and the graphic display section. The second table summary is *InterPro* description table, which displays the non-redundant *InterPro* descriptions for the entire motif or domain hits by the query protein displayed in the first table. If a motif or domain is not mapped to *InterPro*, one can click on the motif link to get its functional description.

The SignalP [24] and TMHMM [25] results are displayed in the third summary table. SignalP predicts if there is a leader peptide, which is the hallmark for secreted proteins and some TM proteins. TMHMM predicts the TM proteins.

PCAS also displays graphically the alignment of each domain or motif relative to the query protein. If exist, leader peptide or TM domains are also indicated.

A link to the annotation data in text format is also provided in the display page (top of Figure 1) so the user can save it for further analysis.

Annotation Coverage

As expected, combination of many computational algorithms increased proteome annotation coverage. As shown in Table 1, for *Arabidopsis*, the overall motif or domain-based annotation coverage in PCAS is about 80%. Not surprisingly, HMMER and RPS have most coverage because of their sensitivity and the larger collection of motif or domains in Pfam/SMART/TIGRFAM and CDD databases. SUPERFAMILY and PRINTS have less coverage. pfscan often produces too many hits. When stringent conditions applied, such as skipping frequently matched patterns and PROSITE entries associated with false SWISSPROT hits [8], pfscan also has low coverage (see table 1). For signal peptide and TM prediction, out of 26032 proteins, SignalP, which consists of SignalP-NN and SignalP-HMM [24], predicts in total 7371 proteins with leader peptide (6932 by SignalP-NN, 4249 by SignalP-HMM; 3810 is the overlap); TMHMM [25] predicts 5971 proteins with TM. For human IPI, PCAS annotation coverage is just over 70%.

Besides Human IPI and *A. thaliana* proteome, Mouse IPI, Rat IPI, *C. elegans*, *D. melanogaster*, *S. cerevisiae*, and *S. pombe* proteomes are also included in PCAS. Zebrafish, Fugu and *C. inestinalis* are being annotated.

We plan to update PCAS at least twice a year limited by the computing resources. We will upgrade PCAS by including new proteome annotation algorithms identified.

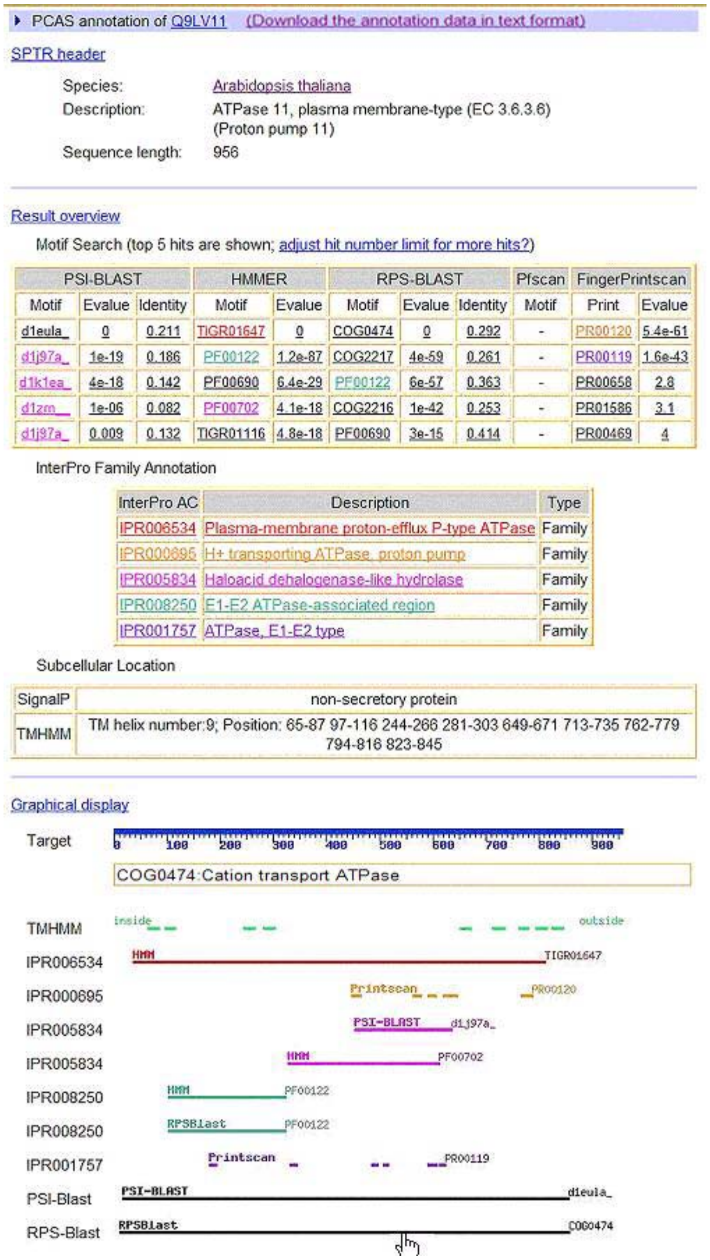
Conclusions

Complementary with the existed proteome annotation efforts, we employed most of the advanced motif and domain based algorithms, to annotate model proteomes. We developed a database and interface system to store and present (query and display) the pre-computed annotation data. We termed this system PCAS (ProteinCentric Annotation System). Comparing with *InterPro*'s daughter profile databases and their respective search methods in *InterProScan*, in PCAS, we also included PSI-BLAST search of protein fold based Superfamily, RPS-BLAST search of NCBI CDD PSSMs, which contains the COGs database and NCBI curated CD database. We employed the internal relations between CDD to map COGs with *InterPro*, and a semi-automatic pipeline to map Superfamily with *InterPro*. This enabled us to integrate most of the motif or domain based annotation data in PCAS through *InterPro* system, and to get rid of the presentation of redundant data from different algorithms. Excluding signal peptide and TM prediction and with this current collection of motif or domain based algorithms, we achieved better annotation coverage by PCAS. Taking human IPI2.18 proteome set as an example, PCAS gave annotation coverage of 70% (Table 1) comparing with *InterPro* at 62% (result came from parsing IPI2.18). PCAS, as most of the computational annotation effort, can thus be used as a discovery platform. We are applying PCAS for novel protein function or novel protein family member discoveries (data not shown).

Availability and requirements

PCAS is available at http://pak.cbi.pku.edu.cn/proteome/gca.php

Figure 1.



▸ PCAS annotation of Q9LV11 (Download the annotation data in text format)

SPTR header

Species:	Arabidopsis thaliana
Description:	ATPase 11, plasma membrane-type (EC 3.6.3.6) (Proton pump 11)
Sequence length:	956

Result overview

Motif Search (top 5 hits are shown; adjust hit number limit for more hits?)

PSI-BLAST			HMMER		RPS-BLAST			Pfscan	FingerPrintscan	
Motif	Evalue	Identity	Motif	Evalue	Motif	Evalue	Identity	Motif	Print	Evalue
d1eula_	0	0.211	TIGR01647	0	COG0474	0	0.292	-	PR00120	5.4e-61
d1j97a_	1e-19	0.186	PF00122	1.2e-87	COG2217	4e-59	0.261	-	PR00119	1.6e-43
d1k1ea_	4e-18	0.142	PF00690	6.4e-29	PF00122	6e-57	0.363	-	PR00658	2.8
d1zrn__	1e-06	0.082	PF00702	4.1e-18	COG2216	1e-42	0.253	-	PR01586	3.1
d1j97a_	0.009	0.132	TIGR01116	4.8e-18	PF00690	3e-15	0.414	-	PR00469	4

InterPro Family Annotation

InterPro AC	Description	Type
IPR006534	Plasma-membrane proton-efflux P-type ATPase	Family
IPR000695	H+ transporting ATPase, proton pump	Family
IPR005834	Haloacid dehalogenase-like hydrolase	Family
IPR008250	E1-E2 ATPase-associated region	Family
IPR001757	ATPase, E1-E2 type	Family

Subcellular Location

SignalP	non-secretory protein
TMHMM	TM helix number:9; Position: 65-87 97-116 244-266 281-303 649-671 713-735 762-779 794-816 823-845

Graphical display

Figure 1

PCAS Display Page. PCAS display of *Arabidopsis* protein **Q9LV11** precomputed annotation data. Query parameters: SUPERFAMILY 1.61 PSI-BLAST E-Value < = 10; Pfam 8.0 hmmpfam E-Value < = 10; CDD 1.62 RPS-BLAST E-Value < = 10, identity > = 0; PRINTS 35.0 FingerPrintscan E-value <= 10; PROSITE pfscan skipping frequently matched patterns; Top hit number < = 5.

Table 1: PCAS Annotation Coverage Five motif or domain based algorithms (see text for respective motif or domain databases) were employed in PCAS to annotate proteomes. The annotation coverage by individual algorithms is listed. [1]pfscan does not score hits; we skipped frequently matching patterns as well as patterns associated with false SWISSPROT hits [8].

Algorithms	**HMMER**	**RPS-BLAST**	**FingerPRINTScan**	**PSI-BLAST**	**Pfscan[1]**
Parameter Setting	Evalue: 0.05	Evalue: 0.01	Evalue: 0.01	Evalue: 0.01	
Arabidopsis (26032):					
Hit number	18491	19499	5621	12950	7090
Annotation coverage	71.0%	74.9%	21.6%	49.7%	27.2%
Total	Hits: 20739	Coverage: 79.7%			
Human IPI (47306):					
Hit number	29601	31156	13949	20887	15607
Annotation coverage	62.6%	65.9%	29.5%	44.2%	33.0%
Total	Hits: 33407	Coverage: 70.6%			
Mouse IPI (35874):					
Hit number	25532	26575	12686	17627	14145
Annotation coverage	71.1%	74.0%	35.3%	49.1%	39.4%
Total	Hits: 28268	Coverage: 78.7%			
Rat IPI (28314):					
Hit number	20673	21350	10190	14898	11014
Annotation coverage	73.0%	75.4%	35.9%	52.6%	38.8%
Total	Hits: 22678	Coverage: 80.1%			
Fly (16807):					
Hit number	11887	12412	4912	8062	5205
Annotation coverage	70.7%	73.8%	29.2%	47.9%	30.9%
Total	Hits: 13332	Coverage: 79.3%			
Worm (21845):					
Hit number	14593	14667	4747	8387	5581
Annotation coverage	66.8%	67.1%	21.7%	38.3%	25.5%
Total	Hits: 16274	Coverage: 74.4%			
S. cerevisiae (6171):					
Hit number	4133	4289	1062	2695	1671
Annotation coverage	66.9%	69.5%	17.2%	43.6%	27.0%
Total	Hits: 4517	Coverage: 73.1%			
S. Pombe (5008):					
Hit number	3714	3854	962	2534	1462
Annotation coverage	74.1%	76.9%	19.2%	50.5%	29.1%
Total	Hits: 4018	Coverage: 80.2%			

List of abbreviations

HMM: Hidden Markov Model. TM: transmembrane region or domain. IPI: International Protein Index. PSSM: Positional Specific Scoring Matrix. CDD: Conserved Domain Database

Authors' contributions

ZY did most of the coding. YYB, ZY, JY, CYJ, GG, YP, LJC participated in collecting, implementing and running of algorithms. JY, YYB and ZY have designed the data model. All authors read and approved the final manuscript.

Acknowledgements

The authors wish to thank Anders Krogh for kindly providing TMHMM, Beijing Institute of Applied Physics and Computational Mathematics, Computing Center of Peking University for computing support. The authors also would like to thank the anonymous reviewers for their helpful comments. This project is supported by Chinese National "863" Program (Grant No. 2001AA231011).

References

1. Falquet L, Pagni M, Bucher P, Hulo N, Sigrist CJ, Hofmann K, Bairoch A: **The PROSITE database, its status in 2002.** *Nucleic Acids Res* 2002, **30:**235-238.
2. Bateman A, Birney E, Cerruti L, Durbin R, Etwiller L, Eddy SR, Griffiths-Jones S, Howe KL, Marshall M, Sonnhammer EL: **The Pfam protein families database.** *Nucleic Acids Res* 2002, **30:**276-280.
3. Letunic I, Goodstadt L, Dickens NJ, Doerks T, Schultz J, Mott R, Ciccarelli F, Copley RR, Ponting CP, Bork P: **Recent improvements to the SMART domain-based sequence annotation resource.** *Nucleic Acids Res* 2002, **30:**242-244.
4. Haft DH, Selengut JD, White O: **The TIGRFAMs database of protein families.** *Nucleic Acids Res* 2003, **31:**371-373.
5. Attwood TK, Blythe MJ, Flower DR, Gaulton A, Mabey JE, Maudling N, McGregor L, Mitchell AL, Moulton G, Paine K, Scordis P: **PRINTS and PRINTS-S shed light on protein ancestry.** *Nucleic Acids Res* 2002, **30:**239-241.
6. Henikoff S, Henikoff JG: **Automated assembly of protein blocks for database searching.** *Nucleic Acids Res* 1991, **19:**6565-6572.
7. Corpet F, Servant F, Gouzy J, Kahn D: **ProDom and ProDom-CG: tools for protein domain analysis and whole genome comparisons.** *Nucleic Acids Res* 2000, **28:**267-269.
8. Gattiker A, Gasteiger E, Bairoch A: **ScanProsite: a reference implementation of a PROSITE scanning tool.** *Applied Bioinformatics* 2002, **1:**107-108.

9. Eddy SR: **Profile hidden Markov models.** *Bioinformatics* 1998, **14:**755-763.
10. Scordis P, Flower DR, Attwood TK: **FingerPRINTScan: Intelligent searching of the PRINTS motif database.** *Bioinformatics* 1999, **15:**799-806.
11. Mulder NJ, Apweiler R, Attwood TK, Bairoch A, Barrell D, Bateman A, Binns D, Biswas M, Bradley P, Bork P, Bucher P, Copley RR, Courcelle E, Das U, Durbin R, Falquet L, Fleischmann W, Griffiths-Jones S, Haft D, Harte N, Hulo N, Kahn D, Kanapin A, Krestyaninova M, Lopez R, Letunic I, Lonsdale D, Silventoinen V, Orchard SE, Pagni M, Peyruc D, Ponting CP, Selengut JD, Servant F, Sigrist CJ, Vaughan R, Zdobnov EM: **The *InterPro* Database, 2003 brings increased coverage and new features.** *Nucleic Acids Res* 2003, **31:**315-318.
12. Zdobnov EM, Apweiler R: ***InterProScan* – an integration platform for the signature-recognition methods in *InterPro*.** *Bioinformatics* 2001, **17:**847-8.
13. Marchler-Bauer A, Panchenko AR, Shoemaker BA, Thiessen PA, Geer LY, Bryant SH: **CDD: a database of conserved domain alignments with links to domain three-dimensional structure.** *Nucleic Acids Res* 2002, **30:**281-283.
14. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Res* 1997, **25:**3389-3402.
15. Mehldau G, Myers G: **A System for Pattern Matching Applications on Biosequences.** *CABIOS* 1993, **9:**299-314.
16. Pesole G, Liuni S, Grillo G, Licciulli F, Larizza A, Makalowski W, Saccone C: **UTRdb and UTRsite: Specializeddatabases of sequences and functional elements of 5' and 3' untranslated regions of eukaryotic mRNAs.** *Nucleic Acids Res* 2000, **28:**193-196.
17. Chen X, Wang LQ, Huang Y, Qiu P, Murgolo NJ, Greene JR, Wu CH, Jiang Y: **IRE_FINDER-Computational search of iron response element in human and mouse UTRs.** *Acta Biochimica et Biophysica Sinica* 2002, **34:**734-747.
18. Jiang Y, Gao G, Fang G, Gustafson EL, Laverty M, Yin Y, Zhang Y, Luo J, Greene JR, Bayne ML, Hedrick JA, Murgolo NJ: **PepPat, a pattern-based oligopeptide homology search method and the identification of a novel Tachykinin-like peptide.** *Mamm Genome* 2003, **14:**341-9.
19. Attwood TK: **The role of pattern databases in sequences analysis.** *Brief Bioinform* 2000, **1:**45-59.
20. Wu CH, Huang H, Yeh LS, Barker WC: **Protein family classification and functional annotation.** *Comput Biol Chem* 2003, **27:**37-47.
21. Murzin AG, Brenner SE, Hubbard T, Chothia C: **SCOP: a structural classification of proteins database for the investigation of sequences and structures.** *Mol Bio* 1995, **247:**536-540.
22. Gough J, Karplus K, Hughey R, Chothia C: **Assignment of Homology to Genome Sequences using a Library of Hidden Markov Models that Represent all Proteins of Known Structure.** *J Mol Biol* 2001, **313:**903-919.
23. Tatusov RL, Koonin EV, Lipman DJ: **A Genomic Perspective on Protein Families.** *Science* 1997, **278:**631-7.
24. Nielsen H, Engelbrecht J, Brunak S, von Heijne G: **Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites.** *Protein Eng* 1997, **10:**1-6.
25. Krogh A, Larsson B, von Heijne G, Sonnhammer EL: **Predicting transmembrane protein topology with a hidden Markov model: Application to complete genomes.** *Journal of Molecular Biology* 2001, **305:**567-580.
26. Gilks WR, Audit B, De Angelis D, Tsoka S, Ouzounis CA: **Modeling the percolation of annotation errors in a database of protein sequences.** *Bioinformatics* 2002, **18:**1641-9.
27. Galperin MY, Koonin EV: **Sources of systematic error in functional annotation of genomes: domain rearrangement, non-orthologous gene displacement and operon disruption.** In *Silico Biol* 1998, **1:**55-67.