410–413 Nucleic Acids Research, 2003, Vol. 31, No. 1
DOI: 10.1093/nar/gkg102

PEP: Predictions for Entire Proteomes

Phil Carter[1,2,*], Jinfeng Liu[1,2,3] and Burkhard Rost[1,2,4]

[1]CUBIC and [2]North East Structural Genomics Consortium (NESG), Department of Biochemistry and Molecular Biophysics, Columbia University, 650 West 168th Street BB217, New York, NY 10032, USA, [3]Department of Pharmacology, Columbia University, 630 West 168th Street, New York, NY 10032, USA and [4]Columbia University Center for Computational Biology and Bioinformatics (C2B2), Russ Berrie Pavilion, 1150 St Nicholas Avenue, New York, NY 10032, USA

Received August 15, 2002; Accepted September 11, 2002

ABSTRACT

PEP is a database of Predictions for Entire Proteomes. The database contains summaries of analyses of protein sequences from a range of organisms representing all three major kingdoms of life: eukaryotes, prokaryotes and archaea. All proteins publicly available for organisms were aligned against SWISS-PROT, TrEMBL and PDB. Additionally, the following annotations are provided: secondary structure, transmembrane helices, coiled coils, regions of low complexity, signal peptides, PROSITE motifs, nuclear localization signals and classes of cellular function. Proteins that contain long regions without regular secondary structure are also identified. We have produced a related database of structural domain-like fragments derived from PEP and clusters based on homology between all fragments. The PEP database, fragments and clusters are distributed freely as a set of flat files and have been integrated into SRS. The PEP group of databases can be accessed from: http://cubic.bioc.columbia.edu/pep.

INTRODUCTION

Large-scale genome sequencing has provided us with the building blocks of living organisms. However, to obtain new insights into physiological and biochemical processes, it is essential to analyse and catalogue the structural and functional features of each individual protein in the genome. We refer to all these proteins as the proteome of an organism. With bioinformatics tools becoming more and more accurate, it is now possible to systematically generate various reliable structural and functional annotations for entire proteomes and make the information easily accessible in different ways. Such predictions for entire proteomes suggest conclusions in context of comparative genomics (1–4) and provide crucial information in the context of structural genomics (4).

DATABASE DESCRIPTION

Design

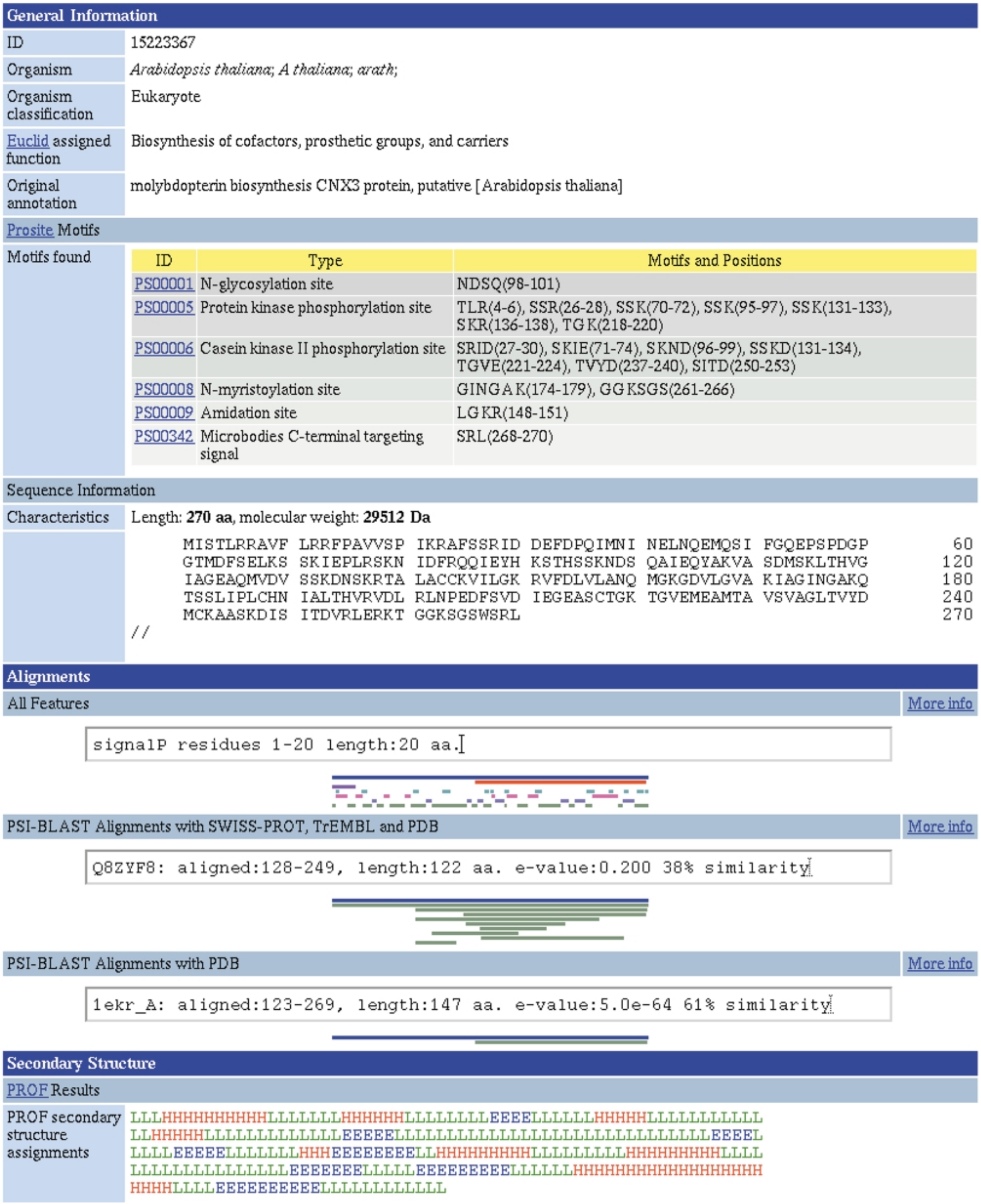
Predictions for Entire Proteomes (PEP) has been created as a generic bioinformatics resource. The objective of predicting features for all constituent peptides of proteomes has been to allow users to data mine proteomes globally, or to retrieve sequences of particular interest and to review predictions on individual sequences. PEP entries constitute the sequences of proteins as given by the Open Reading Frames (ORFs) from sequencing projects. We have dissected the ORFs into putative structural domains or fragments. The fragments in turn have been clustered based upon sequence similarity. The North East Structural Genomics (NESG) consortium (5) is using the fragments and clusters for target selection purposes (http://www.nesg.org).

Content

The PEP database is a summary of analyses for publicly available proteomes (2). All PEP entries were aligned against proteins taken from SWISS-PROT (6), TrEMBL (6) and PDB (7). ORFs were taken from FlyBase (8), WormBase (9) and databases at the NCBI. Protein sequences from each proteome were: (i) aligned against the SWISS-PROT, TrEMBL and PDB using pairwise BLAST (10), PSI-BLAST (11) and the dynamic programming method MaxHom (12); (ii) assigned secondary structure and other sequence based predictions; and (iii) assigned predicted cellular function according to EUCLID (13). The structural and functional features we analysed included:

- coiled-coil regions predicted by COILS (14)
- 3-state secondary structure predicted by PROFsec (15,16)
- percentage relative solvent accessibility predicted by PROFacc (15,16)
- transmembrane helices assigned by PHDhtm (16)
- low sequence complexity regions according to SEG (17)
- long stretches of non-regular secondary structure (NORS) (3)
- presence and location of signal peptide cleavage sites identified by SignalP (18)

[*]To whom correspondence should be addressed. Tel: +1 2123053773; Fax: +1 2123057932; Email: carter@cubic.bioc.columbia.edu



General Information

ID	15223367
Organism	*Arabidopsis thaliana; A thaliana; arath;*
Organism classification	Eukaryote
Euclid assigned function	Biosynthesis of cofactors, prosthetic groups, and carriers
Original annotation	molybdopterin biosynthesis CNX3 protein, putative [Arabidopsis thaliana]

Prosite Motifs

Motifs found

ID	Type	Motifs and Positions
PS00001	N-glycosylation site	NDSQ(98-101)
PS00005	Protein kinase phosphorylation site	TLR(4-6), SSR(26-28), SSK(70-72), SSK(95-97), SSK(131-133), SKR(136-138), TGK(218-220)
PS00006	Casein kinase II phosphorylation site	SRID(27-30), SKIE(71-74), SKND(96-99), SSKD(131-134), TGVE(221-224), TVYD(237-240), SITD(250-253)
PS00008	N-myristoylation site	GINGAK(174-179), GGKSGS(261-266)
PS00009	Amidation site	LGKR(148-151)
PS00342	Microbodies C-terminal targeting signal	SRL(268-270)

Sequence Information

Characteristics: Length: **270 aa**, molecular weight: **29512 Da**

```
      MISTLRRAVF LRRFPAVVSP IKRAFSSRID DEFDPQIMNI NELNQEMQSI FGQEPSPDGP         60
      GTMDFSELKS SKIEPLRSKN IDFRQQIEYH KSTHSSKNDS QAIEQYAKVA SDMSKLTHVG        120
      IAGEAQMVDV SSKDNSKRTA LACCKVILGK RVFDLVLANQ MGKGDVLGVA KIAGINGAKQ        180
      TSSLIPLCHN IALTHVRVDL RLNPEDFSVD IEGEASCTGK TGVEMEAMTA VSVAGLTVYD        240
      MCKAASKDIS ITDVRLERKT GGKSGSWSRL                                         270
//
```

Alignments

All Features — More info

```
signalP residues 1-20 length:20 aa.
```

PSI-BLAST Alignments with SWISS-PROT, TrEMBL and PDB — More info

```
Q8ZYF8: aligned:128-249, length:122 aa. e-value:0.200 38% similarity
```

PSI-BLAST Alignments with PDB — More info

```
1ekr_A: aligned:123-269, length:147 aa. e-value:5.0e-64 61% similarity
```

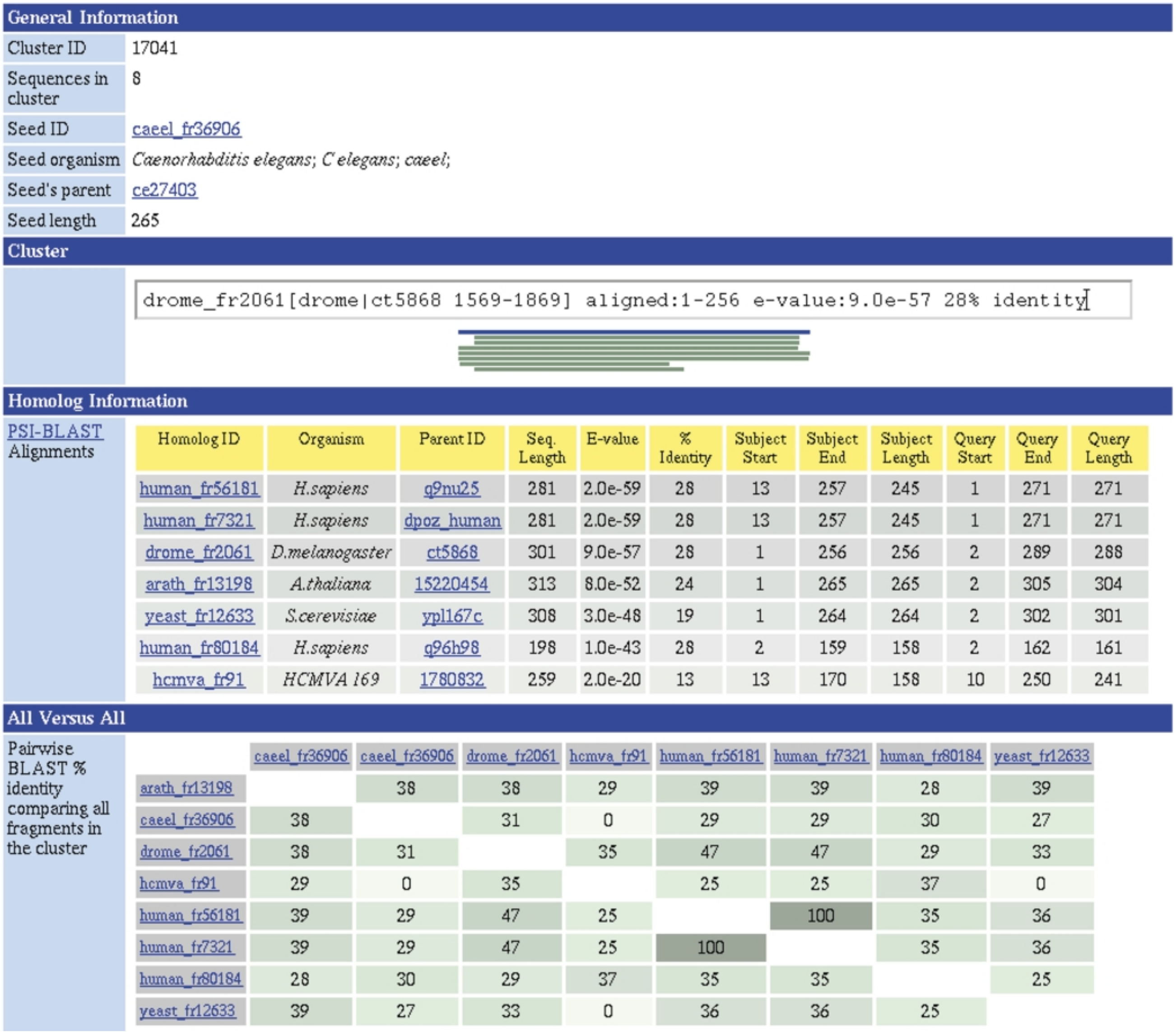
Secondary Structure

PROF Results

PROF secondary structure assignments

```
LLLHHHHHHHHHHLLLLLLLHHHHHHLLLLLLLLEEEELLLLLLHHHHHLLLLLLLLLLL
LLHHHHHLLLLLLLLLLLLLEEEEELLLLLLLLLLLLLLLLLLLLLLLLLLLLLLEEEEL
LLLLEEEEELLLLLLLHHHEEEEEEEELLHHHHHHHHHLLLLLLLLLHHHHHHHHHLLLL
LLLLLLLLLLLLLLLEEEEEEELLLLLEEEEEEEEELLLLLLHHHHHHHHHHHHHHHHHH
HHHHLLLLEEEEEEEEEELLLLLLLLLLLL
```

Figure 1. Screen-dump of a PEP entry. Some general information (organism, sequence length and molecular weight) about the PEP sequence is provided and cellular function as predicted by Euclid. The three graphics are interactive when viewed on the web, and the text above each changes according to the region of the sequence being examined. The first graphic, labelled 'All Features', shows structural and functional features of the sequence and their positions in different colours. In this example, the 270 amino acid sequence is predicted to have a signal peptide 20 residues in length. Also, a long region from residues 123–268 was shown to have homology with a PDB entry. Additionally, helix, beta-sheet and loop regions are indicated. The second and third graphics show the results of PSI-BLAST alignments of the PEP sequence against SWISS-PROT, TrEMBL and PDB databases.



General Information

Cluster ID	17041
Sequences in cluster	8
Seed ID	caeel_fr36906
Seed organism	*Caenorhabditis elegans; C elegans; caeel;*
Seed's parent	ce27403
Seed length	265

Cluster

```
drome_fr2061[drome|ct5868 1569-1869] aligned:1-256 e-value:9.0e-57 28% identity
```

Homolog Information

PSI-BLAST Alignments

Homolog ID	Organism	Parent ID	Seq. Length	E-value	% Identity	Subject Start	Subject End	Subject Length	Query Start	Query End	Query Length
human_fr56181	*H.sapiens*	q9nu25	281	2.0e-59	28	13	257	245	1	271	271
human_fr7321	*H.sapiens*	dpoz_human	281	2.0e-59	28	13	257	245	1	271	271
drome_fr2061	*D.melanogaster*	ct5868	301	9.0e-57	28	1	256	256	2	289	288
arath_fr13198	*A.thaliana*	15220454	313	8.0e-52	24	1	265	265	2	305	304
yeast_fr12633	*S.cerevisiae*	ypl167c	308	3.0e-48	19	1	264	264	2	302	301
human_fr80184	*H.sapiens*	q96h98	198	1.0e-43	28	2	159	158	2	162	161
hcmva_fr91	*HCMVA 169*	1780832	259	2.0e-20	13	13	170	158	10	250	241

All Versus All

Pairwise BLAST % identity comparing all fragments in the cluster

	caeel_fr36906	caeel_fr36906	drome_fr2061	hcmva_fr91	human_fr56181	human_fr7321	human_fr80184	yeast_fr12633
arath_fr13198		38	38	29	39	39	28	39
caeel_fr36906	38		31	0	29	29	30	27
drome_fr2061	38	31		35	47	47	29	33
hcmva_fr91	29	0	35		25	25	37	0
human_fr56181	39	29	47	25		100	35	36
human_fr7321	39	29	47	25	100		35	36
human_fr80184	28	30	29	37	35	35		25
yeast_fr12633	39	27	33	0	36	36	25	

Figure 2. Clustering a structural family. PEP contains clusters of proteins sharing a common structural region corresponding to putative structural domains. Given are the alignments of member sequences against the seed of the cluster produced by PSI-BLAST and results of a pairwise BLAST 'all versus all' comparisons of all the proteins in the cluster.

- PROSITE motifs (19)
- nuclear localization signals (20,21)
- cellular functional classes assigned by EUCLID (13)

An example of a PEP entry is shown in Figure 1.

The structural domain-like fragments have been analysed for the same features i.e. database homologies, sequence based features and cellular function. The fragment results are available as a database named CHOP. These fragments have been clustered using PSI-BLAST with an 'all versus all' sequence similarity comparison to find distinct protein families. The clusters are also available as a database (Fig. 2).

Table 1 shows proteomes we have analysed to date. We will analyse more and add the results to PEP in the future. Currently, we are using a 28 node (58 processor) Dell cluster to perform our predictions.

Availability and interface

The three databases (ORFs, fragments and clusters) are available as flat files and have been integrated into SRS (22). We distribute the full results of the analyses also, although they are quite large in size (gigabytes). The PEP databases can be accessed through the Columbia University Bioinformatics Center (CUBIC) web site: http://cubic.bioc.columbia.edu/pep.

PEP can be searched on many fields (over 40), some examples of which are 'Euclid assigned function', 'number of coiled coil regions', 'length of non-regular secondary structure

Table 1. Excerpt from list of organisms annotated in PEP

Classification	Organism	Number of proteins analysed
Eukaryotes	*Arabidopsis thaliana*	25 542
	Caenorhabditis elegans	20 251
	Drosophila melanogaster	14 304
	Homo sapiens	37 271
	Saccharomyces cerevisiae	6356
Prokaryotes	*Aquifex aeolicus*	1522
	Borrelia burgdorferi	850
	Campylobacter jejuni	1633
	Chlamydia trachomatis	894
	Escherichia coli	4281
	Helicobacter pylori	1564
	Mycoplasma genitalium	470
	Mycoplasma pneumoniae	688
	Neisseria meningitidis	2065
	Rickettsia conorii	1374
	Ureaplasma urealyticum	611
Archaea	*Achaeoglobus fulgidus*	2407
	Aeropyrum pernix K1	2694
	Halobacterium sp. (strain NRC-1)	2058
	Pyrococcus horikoshii	2064
	Sulfolobus solfataricus	2977
Virus	*Human cytomegalovirus* (strain AD169)	202

regions', 'number of alpha-helices', 'number of transmembrane helices' and 'length of signal peptide'. The proteomes can also be searched using a range of bioinformatics tools with their own sequences. The flat files can also be downloaded for local investigation.

ACKNOWLEDGEMENTS

Thanks to Dariusz Przybylski, Rajesh Nair and Kazimierz Wrzeszczynski (Columbia University) for providing preliminary information and programs. Thanks to the SRS team for their software. The work was supported by the grants 1-P50-GM62413-01 and RO1-GM63029-01 from the National Institutes of Health (NIH). Last, not least, thanks to all those who deposit their experimental data in public databases and to those who maintain these databases.

REFERENCES

1. Rost,B. (2002) Did evolution leap to create the protein universe? *Curr. Opin. Struct. Biol.*, **12**, 409–416.
2. Liu,J. and Rost,B. (2001) Comparing function and structure between entire proteomes. *Protein Sci.*, **10**, 1970–1979.
3. Liu,J., Tan,H. and Rost,B. (2002) Loopy proteins appear conserved in evolution. *J. Mol. Biol.*, **322**, 53–64.
4. Liu,J. and Rost,B. (2002) Target space for structural genomics revisited. *Bioinformatics*, **18**, 922–933.
5. Montelione,G.T. (2001) Structural genomics: an approach to the protein folding problem. *Proc. Natl Acad. Sci. USA*, **98**, 13488–13489.
6. Bairoch,A. and Apweiler,R. (2000) The SWISS-PROT protein sequence database and its supplement TFEMBL in 2000. *Nucleic Acids Res.*, **28**, 45–48.
7. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.
8. The Flybase Consortium (2002) The FlyBase database of the Drosophila genome projects and community literature. *Nucleic Acids Res.*, **30**, 106–108.
9. Stein,L., Sternberg,P., Durbin,R., Thierry-Mieg,J. and Spieth,J. (2001) WormBase: network access to the genome and biology of *Caenorhabditis elegans*. *Nucleic Acids Res.*, **29**, 82–86.
10. Altschul,S.F. and Gish,W. (1996) Local alignment statistics. *Methods Enzymol.*, **266**, 460–480.
11. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-Blast: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
12. Sander,C. and Schneider,R. (1991) Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins*, **9**, 56–68.
13. Tamames,J., Ouzounis,C., Casari,G., Sander,C. and Valencia,A. (1998) EUCLID: automatic classification of proteins in functional classes by their database annotations. *Bioinformatics*, **14**, 542–543.
14. Lupas,A. (1996) Prediction and analysis of coiled-coil structures. *Methods Enzymol.*, **266**, 513–525.
15. Rost,B. (2001) Review: protein secondary structure prediction continues to rise. *J. Struct. Biol.*, **134**, 204–218.
16. Rost,B. (1996) PHD: predicting one-dimensional protein structure by profile-based neural networks. *Methods Enzymol.*, **266**, 525–539.
17. Wootton,J.C. and Federhen,S. (1996) Analysis of compositionally biased regions in sequence databases. *Methods Enzymol.*, **266**, 554–571.
18. Nielsen,H., Engelbrecht,J., Brunak,S. and von Heijne,G. (1997) Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Protein Eng.*, **10**, 1–6.
19. Falquet,L., Pagni,M., Bucher,P., Hulo,N., Sigrist,C.J., Hofmann,K. and Bairoch,A. (2002) The PROSITE database, its status in 2002. *Nucleic Acids Res.*, **30**, 235–238.
20. Cokol,M., Nair,R. and Rost,B. (2000) Finding nuclear localization signals. *EMBO Rep.*, **1**, 411–415.
21. Nair,R., Carter,P. and Rost,B. (2003) NLSdb: database of nuclear localization signals. *Nucleic Acids Res.*, **31**, 342–344.
22. Etzold,T. and Argos,P. (1993) Transforming a set of biological flat file libraries to a fast access network. *Comput. Appl. Biosci.*, **9**, 49–57.