Nucleic Acids Research, 2003, Vol. 31, No. 13 **3625–3630**
DOI: 10.1093/nar/gkg545

ELM server: a new resource for investigating short functional sites in modular eukaryotic proteins

Pål Puntervoll, Rune Linding[1], Christine Gemünd[1], Sophie Chabanis-Davidson[1], Morten Mattingsdal, Scott Cameron[2], David M. A. Martin[2], Gabriele Ausiello[3], Barbara Brannetti[3], Anna Costantini[3], Fabrizio Ferrè[3], Vincenza Maselli[3], Allegra Via[3], Gianni Cesareni[3], Francesca Diella[4], Giulio Superti-Furga[4], Lucjan Wyrwicz[5], Chenna Ramu[1], Caroline McGuigan[1], Rambabu Gudavalli[1], Ivica Letunic[1], Peer Bork[1], Leszek Rychlewski[5], Bernhard Küster[4], Manuela Helmer-Citterich[3], William N. Hunter[2], Rein Aasland and Toby J. Gibson[1,*]

Department of Molecular Biology, University of Bergen, Norway, [1]European Molecular Biology Laboratory, Postfach 10.2209, 69012 Heidelberg, Germany, [2]Division of Biological Chemistry and Molecular Microbiology, University of Dundee, Dundee, UK, [3]Centre for Molecular Bioinformatics, Department of Biology, University of Rome Tor Vergata, Rome, Italy, [4]Cellzome AG, Heidelberg, Germany and [5]BioInfoBank Institute, Poznan, Poland

Received February 14, 2003; Revised and Accepted March 26, 2003

ABSTRACT

Multidomain proteins predominate in eukaryotic proteomes. Individual functions assigned to different sequence segments combine to create a complex function for the whole protein. While on-line resources are available for revealing globular domains in sequences, there has hitherto been no comprehensive collection of small functional sites/motifs comparable to the globular domain resources, yet these are as important for the function of multidomain proteins. Short linear peptide motifs are used for cell compartment targeting, protein–protein interaction, regulation by phosphorylation, acetylation, glycosylation and a host of other post-translational modifications. ELM, the Eukaryotic Linear Motif server at http://elm.eu.org/, is a new bioinformatics resource for investigating candidate short non-globular functional motifs in eukaryotic proteins, aiming to fill the void in bioinformatics tools. 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 eliminate false positives. Current filters are for cell compartment, globular domain clash and taxonomic range. In favourable cases, the filters can reduce the number of retained matches by an order of magnitude or more.

INTRODUCTION

The first crystal structure of a protein to be solved, myoglobin, revealed a compact globular structure with regular α-helical elements linked by short irregular loops (1). Because single domain globular proteins are often, though not always, easy to crystallise, for a long time they dominated perception of typical protein structure (although fibrous proteins like collagen were of course well known). Gradually, as protein sequences have accumulated, the monodomain view of protein structure has been replaced by the realisation that most proteins are multidomain, at least in higher eukaryotes. The current champion in size is the giant muscle protein titin at >38 000 residues encompassing some 320 autonomously folded domains (2). Multidomain architectures are usual for transmembrane receptors, signalling proteins, cytoskeletal proteins, chromatin proteins, transcription factors and so forth. There are now several globular protein domain databases accessible on the web, including Pfam (3), SMART (4), PROSITE (5), INTERPRO (6), PRODOM (7) and BLOCKS (8). Using these tools, a user can often get a good overview of the domain architecture of a polypeptide sequence and the functions these domains are likely to perform.

However, there remain protein sequence segments that are difficult to analyse productively. For example, there are often large segments of multidomain proteins that are non-globular, intrinsically lacking the capability to fold into a defined tertiary structure (9–11). Sometimes the function of such regions may be as simple as linkers connecting globular domains and the sequence of amino acids is not important. The structure of

*To whom correspondence should be addressed. Tel: +49 6221387398; Fax: +49 6221387517; Email: toby.gibson@embl.de

The authors wish it to be known that, in their opinion, the first three authors should be regarded as joint First Authors.

yeast RNA polymerase II (12) illustrates this point. Very often, however, these unstructured regions may contain functional sites such as protein interaction sites, cell compartment targeting signals, post-translational modification sites or cleavage sites. These sites are usually short and often reveal themselves in multiple sequence alignments as short patches of conservation, leading to their definition as short sequence motifs. In addition to occurring outside globular domains, some sites, for example, phosphorylation sites, are often found in exposed, flexible loops protruding from within globular domains. These short peptide functional sites are analogous to the linear epitopes of immunology. Considering the abundance of targeting signals and post-translational modification sites, it is reasonable to assume that there are more functional sites than globular domains in a higher eukaryotic proteome.

The PROSITE database has collected a number of linear protein motifs, representing them as regular expression patterns (5). PROSITE patterns have been very useful, but also suffer from severe overprediction problems and more recently the database has emphasised globular domain annotation at the expense of linear motifs. However, the number of known categories of functional sites has burgeoned dramatically in the last few years and it is clear that there are more to be discovered. One only has to think of the huge current research activity into specific methylation and acetylation of histones and chromatin proteins, which erupted after decades of more indirect analyses (13,14). There has been a growing gap in the bioinformatics resources available to researchers for dealing with small functional sites. Indeed, it is impossible for a researcher to find a list of currently known motifs, while going through the literature to retrieve them is impractical without foreknowledge in more areas than any one person will have.

The Eukaryotic Linear Motif (ELM) consortium has established a project to provide a hitherto missing bioinformatics resource for linear motifs. Our aim is to cover the set of functional sites that can be defined by the local peptide sequence, operating essentially independently of protein tertiary structure. The resource suffers from the overprediction problem inherent to small protein motifs, but we are developing context filters such as cell compartment, taxonomy and globular domain clash that can partly reduce the severity of the problem. In this resource, we use the term ELM to denote our bioinformatical representation of a functional site including the sequence motif and its context. ELM is an ongoing project but already provides a working server with >80 motif patterns and access to basic annotation. This manuscript provides an overview of the current status of the ELM resource and an indication of the future directions we hope to take.

ELM RESOURCE ARCHITECTURE

At the core of the ELM resource is a PostgreSQL relational database with 69 tables storing data about linear motifs. Much of this complexity is not yet fully utilised: it anticipates current and future filtering strategies as well as information retrieval by users. The ELM database architecture is beyond the scope of this manuscript and will be presented elsewhere. All data input is by hand curation. Annotating each ELM (our jargon: *Siteseeing*) typically involves extensive literature searches, BLAST runs, multiple alignment of relevant protein families, perusal of SWISS-PROT and other on-line databases and, where practical, discussion with experimentalists from the field. In order to promote interoperability with other bioinformatics resources we use two public annotation standards. Gene Ontology (GO) identifiers are used for cell compartment, molecular function and biological process (15,16) while the NCBI taxonomy database identifiers (17) are used for taxonomic nodes at the apex of phylogenetic groupings in which an ELM occurs. The motif patterns are currently represented as POSIX regular expressions (usable in the Python and Perl languages), analogous to PROSITE, but with a different syntax. For example, the C-terminal peroxisome import signal PTS1 (18) has a consensus sequence of xSKL or KSxL and, allowing for observed redundancy, can be represented as (.[SAPTC][KRH][LMFI]$)|([KRH][SAPTC][NTS][LMFI]$) where $ is the C-terminus.

ELM is primarily developed and deployed with open source software and is hosted on Debian GNU/Linux and secure FreeBSD/OpenBSD systems. Software is developed in Python including some modules from the http://BioPython.org project to retrieve information from SWISS-PROT and PubMed (17). The web interface software uses the CGImodel framework (19). The server output is HTML.

THE ELM SERVER

The public ELM server is at http://elm.eu.org/ and will be mirrored by consortium partners. The scheme in Figure 1 outlines how the server is implemented. Users submit a protein sequence to the server and specify the species and (if known) one or more relevant subcellular compartments. The server reports a list of matching motifs that have been filtered to remove implausible matches. Users should be patient as the turn-around time can be a few minutes while the server accesses several separate resources including the SMART domain server (4). Matched motifs are usually not statistically significant and overprediction will occur despite filtering, hence matches should not be thought to represent true instances of functional sites (unless experimentally verified). Potentially interesting matches might be useful as guides to experiment. The filtered output list has links to the unfiltered results should the user wish to inspect them and also links to retrieve motif annotation from the ELM database.

ELM FILTERS

There is an apparent paradox in sequence motif matching. Pattern methods find many false (but apparently plausible) sequence matches, yet, somehow, these are not recognised by their cognate binding/modification proteins. One obvious reason why a sequence that matches a motif is not a true functional site is that the motif does not fully and accurately represent the functional site. Another reason is that the sequence matches occur in an irrelevant context. They may match to a sequence from a wrong cellular compartment or from a species that does not use this functional site. For these

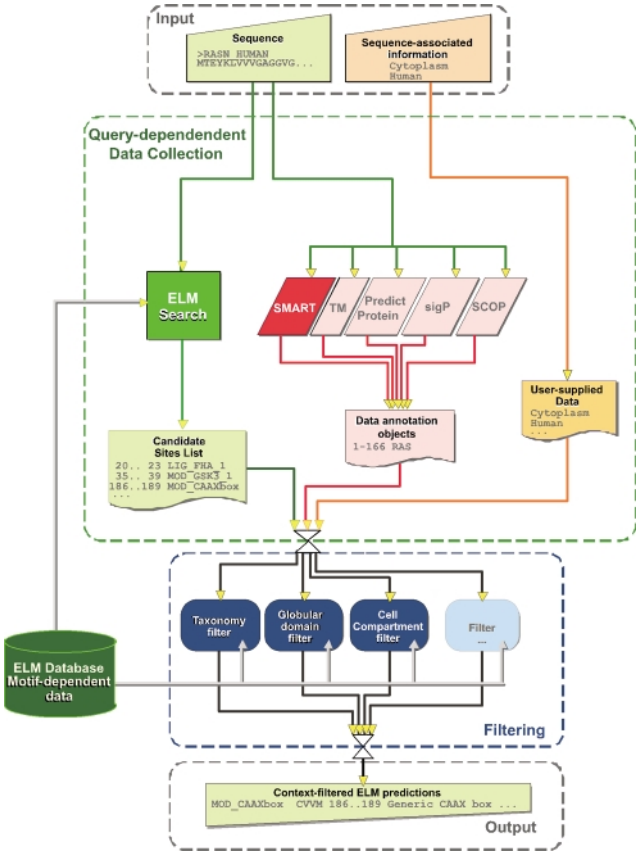

Figure 1. Scheme of the ELM server flowthrough using human RASN as a query. Dashed boxes indicate the four stages from input to result. As the server is further developed, more filters will be added (light blue) requiring more query-dependent data to be retrievable (pink parallelograms).

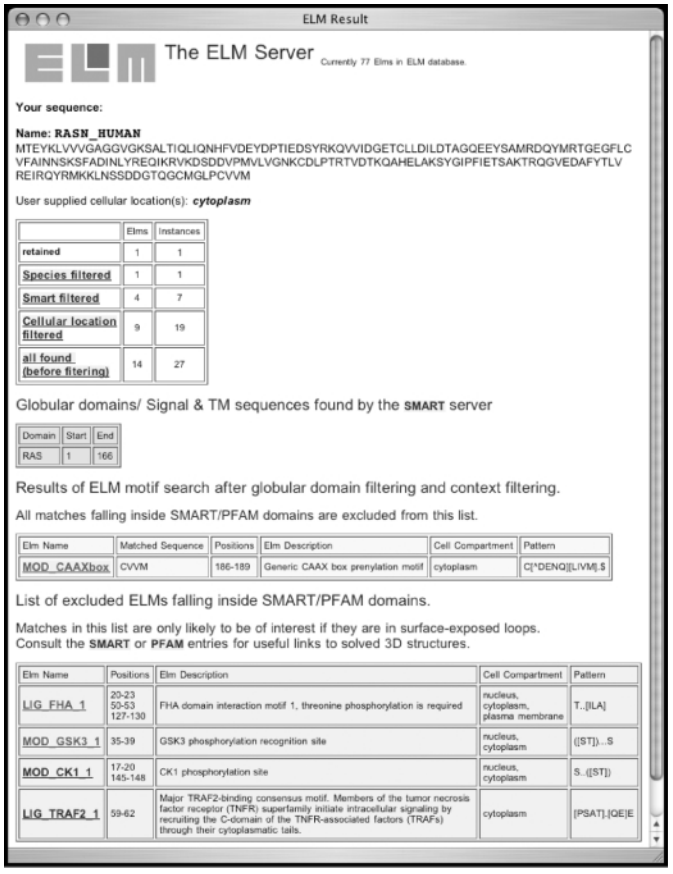

Figure 2. Example of ELM server output using a short sequence, human RASN, as query. The output provides a table summarising the matches and the filtering, a list of globular domains revealed by the SMART server (in this case the RAS domain entry), the list of motif matches that survived filtering (in this case only the C-terminal prenylation site), and finally the list of matches excluded by domain filtering. Hyperlinks to the filtered results as well as to ELM annotation are provided.

cases, it is easy to develop context filters that remove such false positives. Other reasons are less amenable for filter development given current knowledge. For example, tyrosine kinases appear to be non-specific *in vitro* (20,21), yet they may have highly specific substrates *in vivo*. This suggests that their substrates are delivered through adaptor-mediated complexation. We would need to know a lot more about such molecular complexes to deploy them as useful filters. Currently we have three filters installed on the ELM server. These filters are not 100% accurate and may exclude true matches on occasion. The interface provides links to masked matches if the user wishes to retrieve them, but the top level results have been filtered. This approach should encourage users to think critically about ELM server results.

Cell compartment filter

Each ELM will be annotated with GO terms for the set of cell compartments in which it is known to function. For example PTS1 is found on proteins that are targeted to the peroxisomal lumen whereas the NxS/T N-glycosylation site applies to proteins transported out of the cell. The user specifies the compartments in which the query protein functions and all matches for ELMs not found in these compartments will be filtered out.

Globular domain filter

All matches inside globular domains identified with the SMART and Pfam domain databases (3,4) are subtracted. (About 10% of Pfam entries do not in fact correspond to globular domains: at the time of manuscript submission these are part of the filter but we will shortly use flags in Pfam to eliminate them.) Some functional sites seem never to be found inside globular domains, for example, PTS1 or the NR box (LXXLL) (22). However, others, such as phosphorylation sites are frequently in exposed loops of globular domains. Given the limited accuracy of the domain filter, users should consult the unfiltered results too. The domain filter currently acts as a screen. In many cases users will be able to investigate surface accessibility by examination of an available three-dimensional structure or by using a good quality two-dimensional structure prediction (23,24) or perhaps by using a homology modelling server such as SWISSMODEL or the Meta server (25,26). We are working to provide better domain filtering in the future, for example, by using surface accessibility in known structures and annotating known instances of intradomain ELMs.

Table 1. Distribution of selected nuclear ELM matches within SWISS-PROT release 40.41 (121 515 sequence entries)

ELM_ID[a]	Regular expression	Total hits[b]	Taxonomy[c]		Subcellular location[d]			Non-globular[e]
					Nuclear	Non-nuclear	Unknown	
LIG_WRPW	[WFY]RP[WFY].{0,7}$	54	Metazoa	42	34[f]	7	1	34
			Human	10	7	2	1	7
LIG_RBBD	[LI].C.E	6 127	Metazoa	2784	487	1305	992	
			Human	813	185	347	281	87[g]
LIG_NRBOX	[^P]L[^P][^P]LL[^P]	44 902	Metazoa	19 963	2003	11 752	6208	
			Human	6138	775	3641	1722	458
MOD_SUMO	[VILAFP]K.[EDNGP]	255 048	Metazoa	81 329	16 094	37 502	27 733	
			Human	24 319	5968	10 428	7923	4059[h]

[a]LIG_WRPW: ligand motif for transcriptional cofactors; LIG_RBBD: ligand motif for Rb interacting proteins; LIG_NRBOX: ligand motif for nuclear receptors; MOD_SUMO: modification motif for sumoylation.
[b]The total number of regular expression matches. One sequence may have more than one hit.
[c]The taxonomy range for each ELM is given along with the number of matches within that taxonomy range. In addition, the corresponding numbers for *Homo sapiens* are shown.
[d]Subcellular location was evaluated by the SWISS-PROT comment line 'subcellular location.' Nuclear: comment contains word nuclear or nucleus. Non-nuclear: comment does not contain words nuclear or nucleus. Unknown: comment line is missing.
[e]Globularity of the human nuclear sequences with ELM predictions was evaluated by the SMART server (including Pfam domains). All ELMs that are within SMART/Pfam domains were excluded.
[f]All but one of the predicted nuclear LIG_WRPWs are presumptive true positives.
[g]Eleven of 19 experimentally verified instances of LIG_RBBD in human sequences are in this set. Among the missing occurrences are three which are known to reside in globular domains.
[h]Due to the large number of sequences containing predicted MOD_SUMO, 200 randomly chosen sequences were subjected to the SMART/Pfam filtering. The obtained ELM number was scaled to reflect the theoretical number of MOD_SUMOs in nonglobular regions of human nuclear sequences. MOD_SUMO is known to be located in globular domains as well as in nonglobular regions, and some true positives are thus likely to have been filtered out by the crude SMART/Pfam filter.

Taxonomic filtering

Some types of functional site are found in all known eukaryotes, for example, the ER retention signal KDEL is universal (unless there are any eukaryotes that have secondarily lost the endoplasmic reticulum). However, others are restricted to specific eukaryotic taxa. For example, the origin of multicellular animals drove the development of protein export enhancements especially for intercellular communication systems, leading to many novel kinds of functional site. Perhaps most strikingly, the large tyrosine kinase multigene family is found only in Metazoa. Occasionally functional sites may have become secondarily lost in a lineage. An example is PTS2, a second peroxisomal import signal found widely in eukaryotes but absent from the *Caenorhabditis elegans* proteome (27). Each ELM is annotated with one or more NCBI taxonomy node identifiers to indicate its known phylogenetic distribution, for example, the node Metazoa for SH2-, PTB-binding and other phosphotyrosine sites. The user provides the query species and all ELMs that are not assigned to its lineage are filtered out.

Figure 2 shows the ELM server output using the human RASN sequence as a query. Of 77 ELM entries, 14 have matches in the sequence, but 13 are removed by the filters with only the (true) C-terminal prenylation site remaining. This example indicates the potential of logical filters for improving motif searches.

APPLYING ELM

There are two primary purposes motivating the ELM project. One aim is to create a comprehensive database of eukaryotic linear motifs: a knowledge base that is currently missing in biological research. As the resource matures it will become increasingly valuable for data-mining purposes. The second aim is to provide a resource to aid in ELM discovery, furthering the understanding of multidomain proteins. This aim is harder to achieve since the server will provide many false assignments, although this varies according to the sequence information content of the ELMs. We illustrate this by observing the effects of the three currently implemented context filters on four different ELMs occurring in nuclear proteins (Table 1). In the case of WRPW, a motif that occurs at or close to the C-terminus (28), the regular expression alone is highly discriminative; the 54 matches in SWISS-PROT include 33 presumptive true positives. All these are retained after applying the three filters yet only one presumptive false positive remains. At the other extreme is SUMO (29), which has nearly 25 000 matches in the human subset of SWISS-PROT, of which 4059 hits remain after filtering. Since this implies that 3 of 4 of the nuclear proteins have on average ~2.5 sumoylations, this ELM is obviously subject to massive overprediction. Until we are able to provide calibration of ELM results, users can evaluate motif discrimination with the SIRW server (http://sirw.embl.de/), which allows pattern searching of database subsets selected by keyword such as nuclear, cytoplasm or Golgi (30).

Our analysis also shows that the current implementation of the globular domain filter significantly decreases overprediction [for example, by 53% for RBBD (31), see Table 1]. As discussed above, however, some true positives are filtered out since a number of ELMs occur in globular domains. This is the case for RBBD, where three experimentally confirmed sites reside in globular domains (see Table 1, footnote g). This deficiency will be remedied with improved domain filtering.

Table 2. Some specialised resources for motif analysis

Name	Functional sites	URL	PMID
Scansite	Phosphorylation and signaling motifs	scansite.mit.edu	11283593
NetOGlyc	Mucin type GalNAc O-glycosylation sites	www.cbs.dtu.dk/services/NetOGlyc/	9557871
NetNglyc	N-Glycosylation motifs	www.cbs.dtu.dk/services/NetNGlyc/	
PredictNLS	Nuclear localization signals	cubic.bioc.columbia.edu/predictNLS/	11258480
The Sulfinator	Tyrosine sulfation motifs	us.expasy.org/tools/sulfinator/	12050077
NMT	N-terminal N-myristoylation motifs	mendel.imp.univie.ac.at/myristate/SUPLpredictor.htm	11955008
PSORT	Protein sorting signals	psort.nibb.ac.jp/	10829231
TargetP	Protein sorting signals	www.cbs.dtu.dk/services/TargetP/	10891285
SignalP	Cleavage sites and signal/nonsignal peptide prediction	www.cbs.dtu.dk/services/SignalP/	9051728
Big-PI Predictor	GPI modification site	mendel.imp.univie.ac.at/gpi/gpi_server.html	11287675
MITOPROT	Mitochondrial targeting sequences	www.mips.biochem.mpg.de/cgi-bin/proj/medgen/mitofilter	8944766

PMID is the PubMed Identifier for the server publication.

The predictive power of the ELM resource can be enhanced by harnessing it to other data, including experimental results. For example, many protein kinase recognition sites are among those that severely overpredict. If a protein is known not to be phosphorylated, kinase sites can all be ignored, whereas if it is known to be phosphorylated, then the kinase site matches can be targeted for experimental testing. Mass spectrometry can be a useful tool in revealing post-translational modifications. ELM can provide synergism with appropriate experiments and can help in mapping out a research program. In this way, the ELM resource should become increasingly useful to the research community.

OTHER MOTIF RESOURCES

ELM is already the largest collection of linear motifs, followed by PROSITE and Scansite (32). There are other sites that specialise on one or a few motifs for which they may provide better prediction quality than ELM and should be utilised where appropriate. Many functional sites reside in unstructured polypeptide regions and the GlobPlot server (http://globplot.embl.de/) is useful for revealing sequence segments of non-globular character (33), the inverse of the SMART and Pfam domain servers. Some useful motif servers are listed in Table 2 and the ELM and ExPASy servers list more. Also of note are protein interaction databases such as BIND (34) and DIP (35). More informative protein interaction databases that store known instances of linear motifs (36) include MINT (37), Phosphobase (20) and ASC (38). Databases of instances are not directly useful for prediction but provide valuable data-mining resources.

FUTURE DIRECTIONS

The current ELM resource provides basic functionality and there are many ways in which it can be improved. More comprehensive coverage and better motif annotation are planned, including known instances, representative alignments and standardised motif nomenclature (39). In many cases HMM or Profile methods (40) will provide complementary or more sensitive detection with respect to regular expressions and we plan to provide both. We are working to improve filtering logic, especially for globular domains, currently the weakest filter. Other filters, including a surface accessibility filter and a segment flexibility filter, are being developed and will be implemented after successfully passing the benchmarks. Calibration of prediction quality for each ELM is needed for users to assess overprediction likelihoods. The ELM server can be improved with a graphical interface and by performance enhancements that may include GRID technology. We intend to make ELM available for automated proteome analysis pipelines. Last, but not least, we hope that the research community will provide us with useful feedback and help us to improve ELM.

ACKNOWLEDGEMENTS

The ELM consortium is funded by EU grant QLRI-CT-2000-00127.

REFERENCES

1. Kendrew,J.C., Dickerson,R.E., Strandberg,B.E., Hart,R.G. and Davies,D.R. (1960) Structure of myoglobin A three-dimensional Fourier synthesis at 2 Å resolution. *Nature*, **185**, 422–427.
2. Bang,M.L., Centner,T., Fornoff,F., Geach,A.J., Gotthardt,M., McNabb,M., Witt,C.C., Labeit,D., Gregorio,C.C., Granzier,H. and Labeit,S. (2001) The complete gene sequence of titin, expression of an unusual approximately 700-kDa titin isoform, and its interaction with obscurin identify a novel Z-line to I-band linking system. *Circ. Res.*, **89**, 1065–1072.
3. Bateman,A., Birney,E., Cerruti,L., Durbin,R., Etwiller,L., Eddy,S.R., Griffiths-Jones,S., Howe,K.L., Marshall,M. and Sonnhammer,E.L. (2002) The Pfam protein families database. *Nucleic Acids Res.*, **30**, 276–280.
4. Letunic,I., Goodstadt,L., Dickens,N.J., Doerks,T., Schultz,J., Mott,R., Ciccarelli,F., Copley,R.R., Ponting,C.P. and Bork,P. (2002) Recent improvements to the SMART domain-based sequence annotation resource. *Nucleic Acids Res.*, **30**, 242–244.
5. 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.
6. Apweiler,R., Attwood,T.K., Bairoch,A., Bateman,A., Birney,E., Biswas,M., Bucher,P., Cerutti,L., Corpet,F., Croning,M.D. *et al.* (2001) The InterPro database, an integrated documentation resource for protein families, domains and functional sites. *Nucleic Acids Res.*, **29**, 37–40.
7. Corpet,F., Servant,F., Gouzy,J. and Kahn,D. (2000) ProDom and ProDom-CG: tools for protein domain analysis and whole genome comparisons. *Nucleic Acids Res.*, **28**, 267–269.

8. Henikoff,J.G., Greene,E.A., Pietrokovski,S. and Henikoff,S. (2000) Increased coverage of protein families with the blocks database servers. *Nucleic Acids Res.*, **28**, 228–230.
9. Dyson,H.J. and Wright,P.E. (2002) Coupling of folding and binding for unstructured proteins. *Curr. Opin. Struct. Biol.*, **12**, 54–60.
10. Tompa,P. (2002) Intrinsically unstructured proteins. *Trends Biochem. Sci.*, **27**, 527–533.
11. Iakoucheva,L.M., Brown,C.J., Lawson,J.D., Obradovic,Z. and Dunker,A.K. (2002) Intrinsic disorder in cell-signaling and cancer-associated proteins. *J. Mol. Biol.*, **323**, 573–584.
12. Cramer,P., Bushnell,D.A. and Kornberg,R.D. (2001) Structural basis of transcription: RNA polymerase II at 2.8 angstrom resolution. *Science*, **292**, 1863–1876.
13. Strahl,B.D. and Allis,C.D. (2000) The language of covalent histone modifications. *Nature*, **403**, 41–45.
14. Turner,B.M. (2002) Cellular memory and the histone code. *Cell*, **111**, 285–291.
15. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genet.*, **25**, 25–29.
16. Hill,D.P., Blake,J.A., Richardson,J.E. and Ringwald,M. (2002) Extension and integration of the gene ontology (GO): Combining GO vocabularies with external vocabularies. *Genome Res.*, **12**, 1982–1991.
17. Wheeler,D.L., Church,D.M., Lash,A.E., Leipe,D.D., Madden,T.L., Pontius,J.U., Schuler,G.D., Schriml,L.M., Tatusova,T.A., Wagner,L. and Rapp,B.A. (2002) Database resources of the National Center for Biotechnology Information: 2002 update. *Nucleic Acids Res.*, **30**, 13–16.
18. Gatto,G.J.J., Geisbrecht,B.V., Gould,S.J. and Berg,J.M. (2000) Peroxisomal targeting signal-1 recognition by the TPR domains of human PEX5. *Nature Struct. Biol.*, **7**, 1091–1095.
19. Chenna,R. and Gemünd,C. (2000) cgimodel: CGI programming made easy with Python. *Linux J.*, **75**, 142–149.
20. Blom,N., Gammeltoft,S. and Brunak,S. (1999) Sequence and structure-based prediction of eukaryotic protein phosphorylation sites. *J. Mol. Biol.*, **294**, 1351–1362.
21. Kreegipuu,A., Blom,N. and Brunak,S. (1999) PhosphoBase, a database of phosphorylation sites: release 2.0. *Nucleic Acids Res.*, **27**, 237–239.
22. Heery,D.M., Kalkhoven,E., Hoare,S. and Parker,M.G. (1997) A signature motif in transcriptional co-activators mediates binding to nuclear receptors. *Nature*, **387**, 733–736.
23. Cuff,J.A. and Barton,G.J. (2000) Application of multiple sequence alignment profiles to improve protein secondary structure prediction. *Proteins*, **40**, 502–511.
24. Pollastri,G., Przybylski,D., Rost,B. and Baldi,P. (2002) Improving the prediction of protein secondary structure in three and eight classes using recurrent neural networks and profiles. *Proteins*, **47**, 228–235.
25. Guex,N. and Peitsch,M.C. (1997) SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. *Electrophoresis*, **18**, 2714–2723.
26. Bujnicki,J.M., Elofsson,A., Fischer,D. and Rychlewski,L. (2001) Structure prediction meta server. *Bioinformatics*, **17**, 750–751.
27. Motley,A.M., Hettema,E.H., Ketting,R., Plasterk,R. and Tabak,H.F. (2000) Caenorhabditis elegans has a single pathway to target matrix proteins to peroxisomes. *EMBO Rep.*, **1**, 40–46.
28. Paroush,Z., Finley,R.L.J., Kidd,T., Wainwright,S.M., Ingham,P.W., Brent,R. and Ish-Horowicz,D. (1994) Groucho is required for Drosophila neurogenesis, segmentation, and sex determination and interacts directly with hairy-related bHLH proteins. *Cell*, **79**, 805–815.
29. Muller,S., Hoege,C., Pyrowolakis,G. and Jentsch,S. (2001) SUMO, ubiquitin's mysterious cousin. *Nature Rev. Mol. Cell. Biol.*, **2**, 202–210.
30. Ramu,C. (2002) SIRW—a web server for the Simple Indexing and Retrieval System that combines sequence motif searches with keyword searches. *Nucleic Acids Res.*, **31**, 3771–3774.
31. Dahiya,A., Gavin,M.R., Luo,R.X. and Dean,D.C. (2000) Role of the LXCXE binding site in Rb function. *Mol. Cell. Biol.*, **20**, 6799–6805.
32. Yaffe,M.B., Leparc,G.G., Lai,J., Obata,T., Volinia,S. and Cantley,L.C. (2001) A motif-based profile scanning approach for genome-wide prediction of signaling pathways. *Nat. Biotechnol.*, **19**, 348–353.
33. Linding,R., Russell,R.R., Neduva,V. and Gibson,T.J. (2003) globplot—Exploring protein sequences for globularity and disorder. *Nucleic Acids Res.*, **31**, 3701–3708.
34. Bader,G.D., Betel,D. and Hogue,C.W. (2003) BIND: the Biomolecular Interaction Network Database. *Nucleic Acids Res.*, **31**, 248–250.
35. Xenarios,I., Salwinski,L., Duan,X.J., Higney,P., Kim,S.M. and 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.
36. Xenarios,I. and Eisenberg,D. (2001) Protein interaction databases. *Curr. Opin. Biotechnol.*, **12**, 334–339.
37. Zanzoni,A., Montecchi-Palazzi,L., Quondam,M., Ausiello,G., Helmer-Citterich,M. and Cesareni,G. (2002) MINT: a Molecular INTeraction database. *FEBS Lett.*, **513**, 135–140.
38. Facchiano,A.M., Facchiano,A. and Facchiano,F. (2003) Active Sequences Collection (ASC) database: a new tool to assign functions to protein sequences. *Nucleic Acids Res.*, **31**, 379–382.
39. Aasland,R., Abrams,C., Ampe,C., Ball,L.J., Bedford,M.T., Cesareni,G., Gimona,M., Hurley,J.H., Jarchau,T., Lehto,V.P., Lemmon,M.A., Linding,R., Mayer,B.J., Nagai,M., Sudol,M., Walter,U. and Winder,S.J. (2002) Normalization of nomenclature for peptide motifs as ligands of modular protein domains. *FEBS Lett.*, **513**, 141–144.
40. Rychlewski,L., Jaroszewski,L., Li,W. and Godzik,A. (2000) Comparison of sequence profiles. Strategies for structural predictions using sequence information. *Protein Sci.*, **9**, 232–241.