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

Pål Puntervoll, Rune Linding¹, Christine Gemünd¹, Sophie Chabanis-Davidson¹, Morten Mattingsdal, Scott Cameron², David M. A. Martin², Gabriele Ausiello³, Barbara Brannetti³, Anna Costantini³, Fabrizio Ferre³, Vincenza Maselli³, Allegra Via³, Gianni Cesareni³, Francesca Diella⁴, Giulio Superti-Furga⁴, Lucjan Wyrwicz⁵, Chenna Ramu¹, Caroline McGuigan¹, Rambabu Gudavalli¹, Ivica Letunic¹, Peer Bork¹, Leszek Rychlewski⁵, Bernhard Küster⁴, Manuela Helmer-Citterich³, William N. Hunter², Rein Aasland and Toby J. Gibson¹,*

Department of Molecular Biology, University of Bergen, Norway, ¹European Molecular Biology Laboratory, Postfach 10.2209, 69012 Heidelberg, Germany, ²Division of Biological Chemistry and Molecular Microbiology, University of Dundee, Dundee, UK, ³Centre for Molecular Bioinformatics, Department of Biology, University of Rome Tor Vergata, Rome, Italy, ⁴Cellzome AG, Heidelberg, Germany and ⁵BioInfoBank 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 \( \alpha \)-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
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 KStL and, allowing for observed redundancy, can be represented as ([SAPTC][KRH][LMFI][S])([KRH][SAPTC][NTS][LMFI][S]) where S 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
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)

<table>
<thead>
<tr>
<th>ELM_IDa</th>
<th>Regular expression</th>
<th>Total hitsb</th>
<th>Taxonomyc</th>
<th>Subcellular locationc</th>
<th>Non-globulard</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Nuclear</td>
<td>Non-nuclear</td>
<td>Unknown</td>
</tr>
<tr>
<td>LIG_WRPW</td>
<td>[WFY][RP][WFY].[0,7]§</td>
<td>54</td>
<td>42</td>
<td>34f</td>
<td>7</td>
</tr>
<tr>
<td>LIG_RBBD</td>
<td>[LI].C.E</td>
<td>6 127</td>
<td>2784</td>
<td>487</td>
<td>1305</td>
</tr>
<tr>
<td>LIG_NRBOX</td>
<td>[P][P][P][P][IL][P]</td>
<td>44 902</td>
<td>19 963</td>
<td>2003</td>
<td>11 752</td>
</tr>
<tr>
<td>MOD_SUMO</td>
<td>[VILAFP]K.[EDNGP]</td>
<td>255 048</td>
<td>81 329</td>
<td>16 094</td>
<td>37 502</td>
</tr>
</tbody>
</table>

aLIG_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.
bThe total number of regular expression matches. One sequence may have more than one hit.
cThe 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.
dSubcellular 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.

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 furthering the understanding of multidomain proteins. This aim is to provide a resource to aid in ELM discovery, improving domain and motif searches.

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 validation 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.
and we plan to provide both. We are working to improve more sensitive detection with respect to regular expressions methods (40) will provide complementary or HMM or Pro standardised motif nomenclature (39). In many cases planned, including known instances, representative alignments comprehensive coverage and better motif annotation are there are many ways in which it can be improved. More The current ELM resource provides basic functionality and 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.

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

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

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.

ACKNOWLEDGEMENTS

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

REFERENCES

2. Iakoucheva,L.M., Brown,C.J., Lawson,J.D., Obradovic,Z. and
6. Chenna,R. and Gemyu
17. 22. Paroush,Z., Finley,R.L., Kidd,T., Wainwright,S.M., Ingham,P.W.,
targeting signal-1 recognition by the TPR domains of human PEX5. J. Mol. Biol.,
26. 31. Xanarios,I., Salwinski,L., Duan,X.J., Higney,P., Kim,S.M. and
27. 32. 16. Hill,D.P., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M.,
32. 37. Zanzoni,A., Montecchi-Palazzi,L., Quondam,M., Ausiello,G.,
34. 39. Aasland,R., Abrams,C., Ampe,C., Ball,L.I., Bedford,M.T., Cesareni,G.,
of sequence profiles. Strategies for structural predictions using sequence