ELM—the database of eukaryotic linear motifs

Holger Dinkel¹, Sushama Michael¹, Robert J. Weatheritt¹, Norman E. Davey¹, Kim Van Roey¹, Brigitte Altenberg¹, Grischka Toedt¹, Bora Uyar¹, Markus Seiler¹, Aidan Budd¹, Lisa Jödicke¹, Marcel A. Dammert¹, Christian Schroeter¹, Maria Hammer¹, Tobias Schmidt¹, Peter Jehl¹, Caroline McGuigan¹, Magdalena Dymecka², Claudia Chica³, Katja Luck⁴, Allegra Via⁵, Andrew Chatr-aryamontri⁶, Niall Haslam⁷, Gleb Grebnev⁷, Richard J. Edwards⁸, Michel O. Steinmetz⁹, Heike Meiselbach¹⁰, Francesca Diella¹,¹¹ and Toby J. Gibson¹,*

¹Structural and Computational Biology, European Molecular Biology Laboratory, Heidelberg, Germany, ²Laboratory of Bioinformatics and Systems Biology, M. Sklodowska-Curie Cancer Center and Institute of Oncology, WK Roentgena 5, 02-781 Warsaw, Poland, ³Genoscope (CEA – Institut de Génomique), 2 rue Gaston Cremieux CP5706, 91057 Evry, ⁴Group Oncoproteins, Unité CNRS-UDS UMR 7242, Institut de Recherche de l’École de Biotechnologie de Strasbourg, 1, Bd Sébastien Brant, BP 10413, 67412 Illkirch – Cedex, France, ⁵Biocomputing Group, Department of Physics, Sapienza University of Rome, P.le Aldo Moro 5, Rome, Italy, ⁶School of Biological Sciences, University of Edinburgh, Mayfield Road, Edinburgh EH9 3JL, UK, ⁷School of Medicine and Medical Science, University College, Dublin, Ireland, ⁸Centre for Biological Sciences, Institute for Life Sciences, University of Southampton, UK, ⁹Biomolecular Research, Paul Scherrer Institut, CH-5232 Villigen PSI, Switzerland, ¹⁰Bioinformatik, Institut für Biochemie, Friedrich-Alexander-Universität, Fahrstraße 17, 91054 Erlangen-Nürnberg and ¹¹Molecular Health GmbH Belfortstr. 2, 69115 Heidelberg, Germany

Received September 13, 2011; Revised and Accepted October 27, 2011

ABSTRACT

Linear motifs are short, evolutionarily plastic components of regulatory proteins and provide low-affinity interaction interfaces. These compact modules play central roles in mediating every aspect of the regulatory functionality of the cell. They are particularly prominent in mediating cell signaling, controlling protein turnover and directing protein localization. Given their importance, our understanding of motifs is surprisingly limited, largely as a result of the difficulty of discovery, both experimentally and computationally. The Eukaryotic Linear Motif (ELM) resource at http://elm.eu.org provides the biological community with a comprehensive database of known experimentally validated motifs, and an exploratory tool to discover putative linear motifs in user-submitted protein sequences. The current update of the ELM database comprises 1800 annotated motif instances representing 170 distinct functional classes, including approximately 500 novel instances and 24 novel classes. Several older motif class entries have been also revisited, improving annotation and adding novel instances. Furthermore, addition of full-text search capabilities, an enhanced interface and simplified batch download has improved the overall accessibility of the ELM data. The motif discovery portion of the ELM resource has added conservation, and structural attributes have been incorporated to aid users to discriminate biologically relevant motifs from stochastically occurring non-functional instances.

INTRODUCTION

Short linear motifs (SLiMs, LMs or MiniMotifs) are regulatory protein modules characterized by their compact interaction interfaces (the affinity and specificity determining residues are usually encoded between 3 and 11 contiguous amino acids (1)) and their enrichment in natively unstructured, or disordered, regions of proteins (2). As a result of limited intermolecular contacts with their interaction partners, SLiMs bind with relatively
low affinity (in the low-micromolar range), an advanta-
geous attribute for use as transient, conditional and
tunable interactions necessary for many regulatory
processes. Due to the limited number of mutations
necessary for the genesis of a novel motif, SLiMs are
amenable to convergent evolution, functioning as a
driver of network evolution by adding novel interaction
interfaces, and thereby new functionality, to proteins. This
evolutionary plasticity facilitates the rapid proliferation
within a proteome, and as a result, motif use is ubiquitous
in higher eukaryotes.

SLiMs play an important role for many regulatory
processes such as signal transduction, protein trafficking
and post-translational modification (3,4). Their impor-
tance to the correct functionality of the cell is also reflected
by the outcome of motif deregulation. For example,
point mutations in SLiMs have been shown to lead
severe pathologies such as ‘Noonan-like syndrome’ (5),
‘Liddle’s syndrome’ (6) or ‘Retinitis pigmentosa’ (7).
Furthermore, mimicry of linear motifs by viruses to
hijack their hosts’ existing cellular machinery plays an
important role in many viral life cycles (8). However,
despite their obvious importance to eukaryotic cell regu-
lation, our understanding of SLiM biology is relatively
limited, and it has been suggested that, to date, we have
only discovered a small portion of the human motifs (9).

Several resources are devoted to the annotation and/or
detection of SLiMs [Prosite (10), MiniMotifMiner (11)
and Scansite (12)]. Here, we report on the 2012 status
of the Eukaryotic Linear Motif database.

THE ELM RESOURCE

The ELM initiative (http://elm.eu.org) has focused on
gathering, storing and providing information about
short linear motifs since 2003. It was established as the
first manually annotated collection of SLiM classes and
as a tool for discovering linear motif instances in proteins
(13). As it was mainly focused on the eukaryotic sequences, it was termed the Eukaryotic Linear Motif
resource, usually shortened to ELM. The ELM resource
consists of two applications: the ELM database of curated
motif classes and instances, and the motif detection
pipeline to detect putative SLiM instances in query
sequences. In the ELM database, SLiMs are annotated
as ‘ELM classes’, divided into four ‘types’: cleavage
sites (CLV), ligand binding sites (LIG), sites of post-
translational modification (MOD) and subcellular target-
ing sites (TRG) (Table 1). Currently, the ELM database
contains 170 linear motif classes with more than 1800
motif instances linked to more than 1500 literature
references (Table 1). Each class is described by a regular
expression capturing the key specificity and affinity
determining amino acid residues. A regular expression
is a computer-readable term for sequence annotation and is
used by the ELM motif detection pipeline to scan proteins
for putative instances of annotated ELM classes. The
search form for sequence input is shown in Figure 1,
while the results page showing the putative and annotated
instances is illustrated in Figure 2.

The ELM resource is powered by a PostgreSQL
relational database for data storage and a PYTHON
web framework for data retrieval/visualization. The
main tables within the database contain information
about ELM classes, ELM instances, sequences, references,
taxonomy and links to other databases [the database
structure is described in greater detail in (14)].

New ELM classes

Since the last release (14), 24 new ELM classes have been
added to the ELM database (Table 1) and several
more have been updated. One of the newly annotated
motif classes is the AGC kinase docking motif
(LIG_AGCK_PIF), consisting of three distinct classes.
It is present in the non-catalytic C-terminal tail of AGC
kinases that constitute a family of serine/threonine kinases
consisting of 60 members that regulate critical processes,
including cell growth and survival. Deregulation of these
enzymes is a causative factor in different diseases such as
cancer and diabetes. The motif interacts with the PDK1
Interacting Fragment (PIF) pocket in the kinase domain
of AGC kinases. It mediates intramolecular binding to
the PIF pocket, serving as a cis-activating module
together with other regulatory sequences in the C-tail.
Interestingly, in some kinases the motif also acts as a
PDK1 docking site that trans-activates PDK1, which
itself lacks the regulatory C-tail, by interacting with the
PDK1 PIF pocket. PDK1 in turn will phosphorylate and
activate the docked kinase. Other novel classes (Table 2)
include phosphodegrons, which are important mediators
of phosphorylation-dependent protein destruction, and
the LYPxL motif, which is involved in endosomal

Table 1. Summary of data stored in the ELM databasea

<table>
<thead>
<tr>
<th>Number of functional site entries</th>
<th>ELM motif classes</th>
<th>ELM motif instances</th>
<th>Links to PDB structures</th>
<th>GO terms</th>
<th>Pubmed links</th>
</tr>
</thead>
<tbody>
<tr>
<td>Totals</td>
<td>15</td>
<td>170</td>
<td>1840</td>
<td>340</td>
<td>1561</td>
</tr>
<tr>
<td>By category</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LIG</td>
<td>111</td>
<td>Human</td>
<td>1004</td>
<td>Biological process</td>
<td>173</td>
</tr>
<tr>
<td>MOD</td>
<td>30</td>
<td>Mouse</td>
<td>160</td>
<td>From ELM motif</td>
<td>787</td>
</tr>
<tr>
<td>TRG</td>
<td>21</td>
<td>Rat</td>
<td>102</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CLV</td>
<td>8</td>
<td>Fly</td>
<td>67</td>
<td>Cell compartment</td>
<td>74</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Yeast</td>
<td>90</td>
<td>From instance</td>
<td>1071</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Other</td>
<td>417</td>
<td>Molecular function</td>
<td>93</td>
</tr>
</tbody>
</table>

aAs of October 2011.
sorting of membrane proteins but is also implicated in retrovirus budding.

New ELM instances
Annotated ELM instances serve as representative examples of the respective ELM class. They are also invaluable for the computational analysis and classification of motifs (15). Therefore, special emphasis has been put on the curation of more than 500 novel ELM instances (in 40 different classes) by scanning and annotating more than 400 articles. The number of protein databank (PDB) entries annotated have been increased to 195 (Table 1), meaning that for ~10% of all instances there is a 3D

Figure 1. ELM start page. The user can submit a query sequence to the motif detection pipeline either as UniProt accession number or in FASTA format. Filtering criteria such as taxonomic range or cellular compartment should be activated to limit the resulting list of SLiM instances.
protein structure annotated, giving more detailed information about the biological context of the respective motif.

**NEW FEATURES**

The ELM website at http://elm.eu.org can be used in two ways: first, as a front-end to explore the ELM database of curated ELM classes and instances, and second, to run the motif detection pipeline to detect putative SLiM instances in query sequences. Both interfaces have been improved with the most notable changes listed below.

**User interface**

The database user interface, having been stable for many years, has been overhauled and replaced by a novel interface introducing several new features (Figure 1). Up-to-date web technologies have been used to improve the general user experience: the PYTHON framework DJANGO (http://www.djangoproject.com) dynamically creates and serves all HTML pages, while JavaScript was used to make the whole site more interactive and thus improve the user experience. In particular, the ELM detail pages (Figure 3), which hold the most...
Table 2. List of novel ELM classes

<table>
<thead>
<tr>
<th>Identifier</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>LIG_Acitn_WH2_1</td>
<td>Motifs, present in proteins in several repeats, which mediate binding to the hydrophobic cleft created by subdomains 1 and 3 of G-actin</td>
</tr>
<tr>
<td>LIG_Acitn_WH2_2</td>
<td></td>
</tr>
<tr>
<td>LIG_Acitn_RPEL_3</td>
<td></td>
</tr>
<tr>
<td>LIG_AGCK_PIF_1</td>
<td>The AGCK docking motif mediates intramolecular interactions to the PDK1 Interacting Fragment (PIF) pocket, serving as a co-activating module</td>
</tr>
<tr>
<td>LIG_AGCK_PIF_2</td>
<td></td>
</tr>
<tr>
<td>LIG_AGCK_PIF_3</td>
<td></td>
</tr>
<tr>
<td>LIG_BIR_H_1</td>
<td>IAP-binding motifs are found in pro-apoptotic proteins and function in the abrogation of caspase inhibition by inhibitor of apoptosis proteins in apoptotic cells</td>
</tr>
<tr>
<td>LIG_BIR_H_2</td>
<td></td>
</tr>
<tr>
<td>LIG_BIR_H_3</td>
<td></td>
</tr>
<tr>
<td>LIG_BIR_H_4</td>
<td></td>
</tr>
<tr>
<td>LIG_eIF4E_1</td>
<td>Motif binding to the dorsal surface of eIF4E</td>
</tr>
<tr>
<td>LIG_eIF4E_2</td>
<td></td>
</tr>
<tr>
<td>LIG_EVH1_3</td>
<td>A proline-rich motif binding to EVH1/WH1 domains of WASP and N-WASP proteins</td>
</tr>
<tr>
<td>LIG_HCF-1_HBM_1</td>
<td>The DHxY Host Cell Factor-1 binding motif interacts with the N-terminal kelch propeller domain of the cell cycle regulator HCF-1</td>
</tr>
<tr>
<td>LIG_Integrin_isoDGR_1</td>
<td>Present in proteins of extracellular matrix which upon deamidation forms biologically active isoDGR motif which binds to various members of integrin family</td>
</tr>
<tr>
<td>LIG_LYPXl_L_2</td>
<td>The LYPxL motif binds the V-domain of Alix, a protein involved in endosomal sorting</td>
</tr>
<tr>
<td>LIG_LYPXl_S_1</td>
<td></td>
</tr>
<tr>
<td>LIG_PAM2_1</td>
<td>Peptide ligand motif that directly interacts with the MLLE/PABC domain found in poly(A) binding proteins and HYD E3 ubiquitin ligases</td>
</tr>
<tr>
<td>LIG_PIKK_1</td>
<td>Motif located in the C terminus of Nbs1 and its homologous interacting with PIKK family members</td>
</tr>
<tr>
<td>LIG_Rb_pAAbgroove_1</td>
<td>The LxxLFD motif binds in a deep groove between pocket A and pocket B of the Retinoblastoma protein</td>
</tr>
<tr>
<td>LIG_SCF_FBW7_1</td>
<td>The TPxxS phospho-dependent degron binds the FBW7 F box proteins of the SCF (Skp1-Cullin-Fbox) complex</td>
</tr>
<tr>
<td>LIG_SCF_FBW7_2</td>
<td></td>
</tr>
<tr>
<td>LIG_SPAK-OSR1_1</td>
<td>SPARK-OSR1 kinase binding motif acts as a docking site which aids the interaction with their binding partners including the upstream activators and the phosphorylated substrates</td>
</tr>
</tbody>
</table>

As of October 2011.

important information about each ELM class including references, regular expression, taxonomic distribution and gene ontology terms (Table 3), have been updated by annotating the protein domain interacting with the respective motif. Where available, a 3D model of representative protein databank structures of linear motif interactions was added to the ELM detail page (Figure 3, top right).

To cope with the increasing amount of annotated classes as well as instances, a novel query interface was introduced to assist the user in finding information of interest. The ELM browser (Figure 4) now features a search interface for free text search. In addition, the search results can also be filtered and reordered using buttons (Figure 4, left side) and table headers, respectively, and be downloaded as tab-separated values (TSV).

Further, improvements to the ELM database include revising the experimental methods used for annotation by using a standardized methods vocabulary [in sync with PSI-MI ontology (16,17)].

A candidate page has been introduced to display novel ELM classes that have not yet been annotated in detail or are currently undergoing annotation. We invite researchers to send us their feedback and expert opinion on these classes and to contribute novel motif classes that will be added to the candidate page and ultimately be turned into full ELM classes (Figure 5). Minimum requirements are at least one literature reference as well as a short description. In addition, a draft regular expression or a 3D structure showing the relevant interaction would also be helpful. Currently, the number of possible ELM classes on this candidate list (awaiting further annotation) exceeds the number of completely annotated classes, indicating the great demand for further annotation.

Graphical representation of sequence search

The ELM motif detection pipeline scans protein sequences for matches to the regular expressions of annotated ELM classes (Figure 2). The query output combines these putative instances with information from the database (annotated ELM instances) as well as predictions from different algorithms/filters. The ELM resource employs a structural filter (18) to highlight and mask secondary structure elements, as well as SMART (19) to detect protein domains. Furthermore, an additional disorder prediction algorithm (IUPred) (20) has been included to predict ordered/disordered regions within the protein. IUPred uses a cutoff of 0.5 to classify a sequence region as either structured or disordered, with values above this threshold corresponding to disorder, highlighted in green background and lower values indicating structured regions, displayed in red background in the output graph. Disorder and domain information is combined by
The conservation of linear motifs can help in assessing the functional relevance of putative instances, with functional instances showing higher overall sequence conservation than non-functional ones (21). Therefore, sequence conservation of the query protein is calculated using a tree-based conservation scoring method (22) and highlighted in the graphical output. Here, lighter shades of blue represent low conservation while dark blue shading corresponds to high-sequence conservation. The actual conservation score can be inspected by moving the mouse over the respective ELM instance (Figure 2).

The functionality of linear motifs can be modulated by modifications such as phosphorylation (23,24). To enable the user to investigate phosphorylation data in the context of putative linear motif instances, phosphorylation annotations from the Phospho.ELM resource (25) have been added to the graphical output (Figure 2, top row).

Table 3. Main cellular compartments used in ELM annotation

<table>
<thead>
<tr>
<th>Count</th>
<th>GO Id</th>
<th>GO term</th>
</tr>
</thead>
<tbody>
<tr>
<td>98</td>
<td>GO:0005829</td>
<td>Cytosol</td>
</tr>
<tr>
<td>69</td>
<td>GO:0005634</td>
<td>Nucleus</td>
</tr>
<tr>
<td>17</td>
<td>GO:0005576</td>
<td>Extracellular</td>
</tr>
<tr>
<td>12</td>
<td>GO:0005794</td>
<td>Golgi apparatus</td>
</tr>
<tr>
<td>10</td>
<td>GO:0005886</td>
<td>Plasma membrane</td>
</tr>
<tr>
<td>9</td>
<td>GO:0009998</td>
<td>Internal side of plasma membrane</td>
</tr>
<tr>
<td>9</td>
<td>GO:0005783</td>
<td>Endoplasmic reticulum</td>
</tr>
<tr>
<td>6</td>
<td>GO:0005739</td>
<td>Mitochondrion</td>
</tr>
<tr>
<td>5</td>
<td>GO:0005643</td>
<td>Nuclear pore</td>
</tr>
<tr>
<td>5</td>
<td>GO:0045334</td>
<td>Clathrin-coated endocytic vesicle</td>
</tr>
</tbody>
</table>
The phosphorylated residues are highlighted in different colors (serine: green, threonine: blue, tyrosine: red); each phosphorylation site is linked to a page showing detailed information about the respective modification site from the manually curated data set of the Phospho.ELM resource.

**VIRAL INSTANCES**

The importance of the short linear motifs in virus–host interactions makes the ELM resource an important tool for the viral research community. For example, Cruz et al. (26) analyzed a protein phosphatase 1 (PP1) docking motif in ‘protein 7’ of transmissible gastroenteritis virus using the ELM class LIG_PP1. This conserved sequence motif mediates binding to the PP1 catalytic subunit, a key regulator of the cellular antiviral defense mechanisms, and is also found in other viral proteomes, suggesting that it might be a recurring strategy to counteract the hosts’ defense against RNA viruses by dephosphorylating eukaryotic translation initiation factor 2α and ultimately ribonuclease L.

To reflect our increasing awareness of viral motifs (8), special focus has been attributed to the annotation of viral instances in the ELM database: in the latest release, more than 200 novel ELM instances found in 84 different viral taxons have been added. The notion of viruses abusing existing SLiMs in their hosts is demonstrated by viral instances being annotated alongside instances in their hosts’ proteins. For example, the ELM class LIG_PDPZ_Class_1 contains 12 instances in human proteins but has recently been expanded with 5 instances from 5 different human pathogenic virus proteins.

**Figure 4.** ELM instances browse page. A full-text search (here, search term used was ‘AP2’, filtering for ‘true positive’ instances in taxon ‘*Homo sapiens*’, yielding 58 instances) assists in finding annotated instances. A search can be restricted to a particular taxonomy or instance logic (top) or ELM class type (buttons on the left). The list can also be exported to TSV or FASTA format for further processing.
of these interactions results in myofibrillar myopathies (32). Additionally, ELM annotations can contribute to high-throughput screenings (33) as well as development of novel algorithms (34–36), methods (37) and databases (38). Furthermore, the highly curated data of the ELM resource are used as a benchmarking data set to evaluate the accuracy of prediction algorithms (21,39,40).

For any such analysis, the user should be aware that many matches to ELM regular expressions are false positives. Before conducting experiments based on ELM results, it is strongly advisable to check if a motif match is conserved, exposed in a cell compartment in which the motif is known to be functional. The ELM resource applies several filters to provide the user with such information that should ideally also be supported by the experimental evidence.

SUMMARY

The importance of SLiMs is highlighted by the growing number of instances with relevance to diseases or viruses. Yet, despite their importance and abundance, our understanding of linear motifs is still limited. This is mainly owing to the fact that they are still quite difficult to predict computationally and to investigate experimentally (3,41,42). By better understanding the biology of linear motifs, we hope to increase our insight into diseases and viruses (and vice versa). The ELM resource tries to aid the researcher in the search for putative SLiM instances by providing a feature-rich toolset for sequence analysis. Consequently, with the aforementioned additions and changes, we hope that the ELM resource continues to be a valuable asset to the community.

ACKNOWLEDGEMENTS

The authors would like to thank the users of the ELM resource as well as all colleagues, contributors and annotators of the ELM resource.

FUNDING

EMBL international PhD program (to R.J.W.); EMBL Interdisciplinary PostDoc fellowship (EIPoD to N.E.D.); NGFN framework by the Federal Government Department of Education and Science [FKZ01GS0862 (DiGtoP) to M.S. and M.H.]; European Community’s Seventh Framework Programme FP7/2009 (SysCilia) (241955 to G.T.) and (SysBoSS) (242129 to K.V.R.); Polish Ministry of Science and Higher Education within Iuventus Plus project (IP2010-0483-70 to M.D.); Biotechnology and Biological Sciences Research Council (BB/F010486/1 to A.C.); Région Alsace and Collège Doctoral Européen (to K.L.); Science Foundation Ireland (08/N1.1/B1864 to G.G.); BBSRC New Investigator Award (BB/1006230/1 to R.J.E.); German Research Foundation (SFB796 Project A2 to H.M.); grants from the Swiss National Science Foundation (to M.O.S.). Funding for open access charge: EMBL.

Conflict of interest statement. None declared.
REFERENCES


