SLiMSearch: A Webserver for Finding Novel Occurrences of Short Linear Motifs in Proteins, Incorporating Sequence Context

Norman E. Davey[1], Niall J. Haslam[2],
Denis C. Shields[2], and Richard J. Edwards[3]

[1] Structural and Computational Biology Unit,
European Molecular Biology Laboratory, 69117 Heidelberg, Germany
[2] School of Medicine and Medical Sciences, UCD Complex and Adaptive Systems Laboratory & UCD Conway Institute of Biomolecular and Biomedical Sciences,
University College Dublin, Dublin, Ireland
[3] School of Biological Sciences, University of Southampton, Southampton, UK
davey@embl.de,
{niall.haslam,denis.shields}@ucd.ie,
r.edwards@southampton.ac.uk

Abstract. Short, linear motifs (SLiMs) play a critical role in many biological processes. The SLiMSearch (Short, Linear Motif Search) webserver is a flexible tool that enables researchers to identify novel occurrences of pre-defined SLiMs in sets of proteins. Numerous masking options give the user great control over the contextual information to be included in the analyses, including evolutionary filtering and protein structural disorder. User-friendly output and visualizations of motif context allow the user to quickly gain insight into the validity of a putatively functional motif occurrence. Users can search motifs against the human proteome, or submit their own datasets of UniProt proteins, in which case motif support within the dataset is statistically assessed for over- and under-representation, accounting for evolutionary relationships between input proteins. SLiMSearch is freely available as open source Python modules and all webserver results are available for download. The SLiMSearch server is available at: http://bioware.ucd.ie/slimsearch.html.

Keywords: short linear motif, motif discovery, minimotif, elm.

1 Introduction

The purpose of the SLiMSearch (Short, Linear Motif Search) webserver is to allow researchers to identify novel occurrences of pre-defined Short Linear Motifs (SLiMs) in a set of sequences. SLiMs, also referred to as linear motifs or minimotifs, are functional microdomains that play a central role in many diverse biological pathways [1]. SLiM-mediated biological processes include post-translational modification (including cleavage), subcellular localization, and ligand binding [2]. SLiMs are typically less than ten amino acids long and have less than five defined positions, many of which will be “degenerate” and incorporate some degree of flexibility in

T.M.H. Dijkstra et al. (Eds.): PRIB 2010, LNBI 6282, pp. 50–61, 2010.

terms of the amino acid at that position. Their length and degeneracy gives them an evolutionary plasticity which is unavailable to domains meaning that they will often evolve convergently, adding new functionality to proteins [1]. SLiMs hold great promise as future therapeutic targets, which makes their discovery of great interest [3-4].

Once a SLiM has been defined, finding matches in a given set of protein sequences is a fairly trivial task. Finding biological motifs is a standard pattern recognition task in bioinformatics. Several web-based methods to discover novel instances of known SLiMs are available, including ELM [2], MnM [5], SIRW [6] ScanProsite [7] and QuasiMotifFinder [8], which generally utilize databases of known motif patterns to search query protein sequences supplied by the user. Whilst finding matches is trivial, however, interpreting their biological significance is far from easy. The small, degenerate nature of SLiMs makes stochastic occurrences of motifs common; distinguishing real occurrences from the background of random motif hits remains the greatest challenge in *a priori* motif discovery. One approach is to simply filter out motifs that are likely to occur numerous times by chance – ScanProsite [7], for example, has an option to "Exclude motifs with a high probability of occurrence", while QuasiMotifFinder [8] uses the background occurrence of motifs in PfamA families [9] to assess the significance of hits. These strategies work well for longer, family descriptor motifs (such as are found in the Prosite database [10] used by both ScanProsite and QuasiMotifFinder) but are not so useful for SLiMs because of their tendency to occur by chance. Instead, additional contextual information such as sequence conservation [5, 8, 11-12], structural context [5, 13] or even biological keywords [6] can be used to assess the likelihood of true functional significance for putatively functional sites.

Most motif search tools rely on pre-existing motif libraries, such as ELM [2], MnM [5] or Prosite [10]. Those that permit users to define their own motifs, such as ScanProsite [7], are generally lacking the contextual information required to aid functional inference. Recent developments in *de novo* motif discovery has given rise to a number of tools that are capable of predicting entirely novel SLiMs from sets of protein sequences (*e.g.* PRATT [14], MEME [15], Dilimot [16], SLiMDisc [17] and SLiMFinder [18]). Although SLiMFinder [18] estimates the statistical significance of returned motif predictions, correcting for biases introduced by evolutionary relationships within the data, assessing the *biological* significance of predicted SLiMs remains challenging. On approach is to compare candidate SLiMs to existing motif libraries to identify similarities to previously known motifs [19].When a genuinely novel motif is predicted, however, knowledge of existing motifs is of limited use. Instead, it is useful to be able to establish the background distribution of occurrences of the novel motif, utilizing contextual information to help screen out the inevitable spurious chance matches.

We recently made our powerful *de novo* SLiM discovery tool, SLiMFinder [18], available as a webserver [20]. To aid interpretation of SLiMFinder results, we have made a new tool available, SLiMSearch, which allows users to search protein datasets with user-defined motifs, including motif prediction output from SLiMFinder. SLiMSearch utilizes the same sequence context assessment as SLiMFinder, enabling results to be masked or ranked based on the important biological indicators of sequence conservation and structural disorder [12, 21]. SLiMSearch also features the same SLiMChance algorithm for assessing statistical over-representation of SLiM occurrences, correcting for biases introduced by evolutionary relationships within the data. SLiMSearch is open source and freely available for download. For ease of use, the main SLiMSearch features have been made available as a webserver, which enables the user to search proteins for occurrences of user-specified motifs. Motifs can be searched against small custom datasets of proteins from UniProt [22]. Alternatively, searches can be performed against the whole human proteome, or defined subsets of it. Underlying methods, results formats and visualizations are fully compatible with our existing SLiM analysis webservers, SLiMDisc [23], CompariMotif [19] and SLiMFinder [20], providing a suite of integrated tools for analyzing these biologically important sequence features.

2 The SLiMSearch Algorithm

SLiMSearch performs its motif finding in three phases: (1) Input sequences are read and masked; (2) Motifs are searched against masked sequences using standard regular expression searches; (3) Motif statistics are calculated for identified motif occurrences. If desired, input sequences, input motifs and motif occurrences can be filtered based on attributes such as length, number of positions, motif conservation *etc.* SLiMs have a tendency to occur in disordered regions of proteins [24] and IUPred [21] protein disorder predictions can be used for input masking or ranking/filtering results as described further below. Conservation scoring uses the Relative Local Conservation (RLC) score introduced by Davey *et al.* [12] as implemented in SLiMFinder [20]. Conservation scoring can use pre-generated alignments or construct alignments of predicted orthology using GOPHER [23], which estimates evolutionary relationships using BLAST [25] to identify the closest-related orthologue in each species in the chosen search database. Each putative orthologue retained is: (a) more closely related to the query than any other protein from the same species; (b) related to the query through a predicted speciation event, not a duplication event.

2.1 SLiMChance Calculations of Significance

SLiMSearch utilizes a variation of the SLiMChance algorithm from SLiMFinder [18], which is based on the binomial statistics introduced by ASSET [26] and calculates the *a priori* probability of observing each motif in each sequence using the (masked) amino acid frequencies of input sequences. Observed support is then compared to expectation at two levels: (1) the total number of occurrences in all sequences; (2) the number of individual sequences returning the motif. This enables different questions to be asked of different data types. SLiMChance has an important extension over the statistics used by ASSET, and homologous proteins are optionally weighted (as in SLiMDisc [17] and SLiMFinder [18]) to account for the dependencies introduced into the probabilistic framework by homologous proteins; in this case, SLiMSearch will also assess these weighted support values. Whereas SLiMFinder is explicitly using *over*-representation to identify motifs, it is also of potential interest to see if a given motif has been avoided in a given dataset and is *under*-represented versus random expectation. The SLiMSearch implementation of SLiMChance therefore features an additional extension where the cumulative binomial probability is used to estimate the probability of seeing by chance the observed support *or less* in addition to the observed support *or more*.

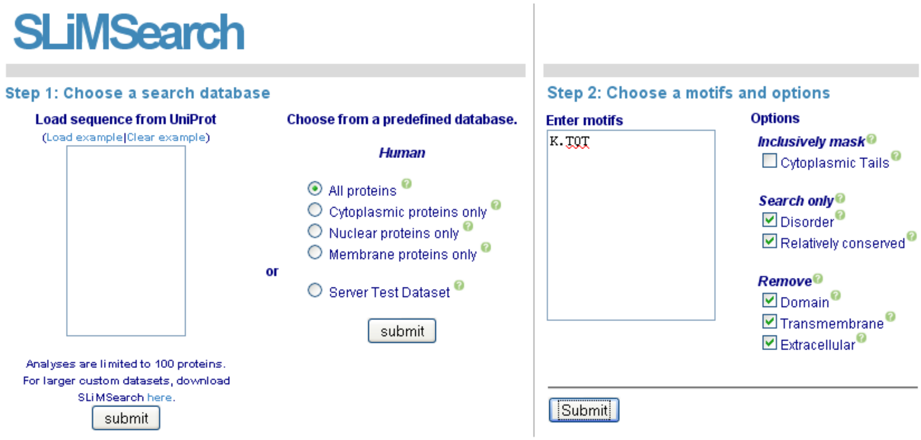
3 The SLiMSearch Webserver

The SLiMSearch server is available at: http://bioware.ucd.ie/slimsearch.html. The purpose of the webserver is to allow researchers to identify novel occurrences of predefined Short Linear Motifs (SLiMs) in a set of protein sequences. Sequences are first masked according to user specifications before motif occurrences are identified using standard regular expression searches. The SLiMChance algorithm then estimates statistical significance of over- or under-representation of each motif searched. In addition to summary results for each motif, interactive output permits easy exploration and visualization of individual motif occurrences. The context of each SLiM occurrence is then calculated in terms of protein disorder and evolutionary conservation to help the user gain insight into the validity of a putatively functional motif occurrence. The webserver is powered by the same code as the standalone version of SLiMSearch, which can be downloaded from the server. The main features of the webserver are described in more detail in the following sections.

3.1 Input

As input, SLiMSearch needs a set of protein sequences and a set of motif definitions, which are selected by the user in turn (Fig. 1). Whereas the standalone SLiMSearch program allows searching of any protein sequences, the webserver restricts the user to using UniProt sequences [22]. This is because the server relies on pre-computed alignments to keep run times down. Using UniProt downloads also allows all the masking options to be utilized (*e.g.* sequence features). The user is presented with a choice of two main input types (Fig. 1): (1) a chosen set of up to 100 UniProt entries can be downloaded for analysis; (2) the user can select from a series of predefined protein datasets. Currently, the human proteome from SwissProt [22] is available, along with three subsets defined by their subcellular localization annotation: cytoplasmic proteins, nuclear proteins and transmembrane proteins. Future server releases will expand this to other species. When searching these large proteome datasets, the evolutionary filtering [18] is switched off. To search different datasets, including datasets over 100 proteins with evolutionary filtering, users are encouraged to download and install a local version of SLiMSearch.

Once a dataset has been selected, the user must input a set of motifs to search (Fig. 1). The SLiMSearch server takes a list of motifs, typed or pasted directly into the text box. Motifs themselves are constructed from a number of regular expression elements, which are mostly standard but with a couple of additional elements to represent “3of5” motifs [27] (Table 1). SLiMSearch accepts the same input formats as CompariMotif [19], including a plain list of regular expressions and output from SLiMDisc [23] or SLiMFinder [20]. Because the focus of SLiMSearch is *short* linear motifs, the maximum number of consecutive wildcards allowed by the server is nine. Motifs must have at least *two* defined (*i.e.* non-wildcard) positions.



SLiMSearch

Step 1: Choose a search database

Load sequence from UniProt

(Load example|Clear example)

Analyses are limited to 100 proteins. For larger custom datasets, download SLiMSearch here.

submit

or

Choose from a predefined database.

Human

● All proteins
○ Cytoplasmic proteins only
○ Nuclear proteins only
○ Membrane proteins only

○ Server Test Dataset

submit

Step 2: Choose a motifs and options

Enter motifs

```
K.TQT
```

Options

Inclusively mask
□ Cytoplasmic Tails

Search only
☑ Disorder
☑ Relatively conserved

Remove
☑ Domain
☑ Transmembrane
☑ Extracellular

Submit

Fig. 1. SLiMSearch input options pages. Users must first either select a predefined human protein dataset, or enter a list of up to 100 UniProt IDs for a custom dataset. Clicking "submit" will then progress to Step 2, in which users enter a list of motifs for searching and set any masking options.

Table 1. Regular expression elements recognized by SLiMSearch

Element	Description
A	Single fixed amino acid.
[AB]	Ambiguity, **A** or **B**. Any number of options may be given, *e.g.* **[ABC]** = **A** or **B** or **C**.
<R:m:n>	At least **m** of a stretch of **n** residues must match **R**, where **R** is one of the above regular expression elements (single or ambiguity).
<R:m:n:B>	Exactly **m** of a stretch of **n** residues must match **R** and the rest must match **B**, where **R** and **B** are each one of the above regular expression elements (single or ambiguity). *E.g.* **<F:1:2:[DE]>** will match **[DE]F**, or **F[DE]**.
[^A]	Not **A**.
X or **.**	Wildcard positions (any amino acid).
.{m,n}	At least **m** and up to **n** wildcards.
R{n}	**n** repetitions of **R**, where **R** is any of the above regular expression elements.
^	Beginning of sequence
$	End of sequence
(R|S)	Match **R** or **S**, which are both themselves recognizable regular expressions. These motifs are not currently supported by the SLiMChance statistics and, as such, any motifs in this format with be first split into variants, *e.g.* **(R|S)PP** would be split into **RPP** and **SPP** and each searched separately.

3.2 Masking Options

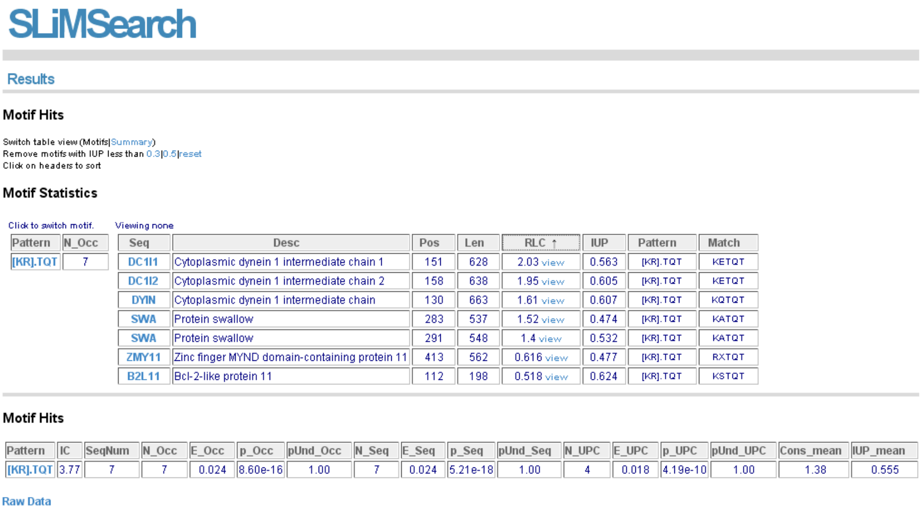
The standalone SLiMSearch program features all the input masking options of SLiMFinder [18]. For simplicity, these have been pared down for the webserver to three sets of masking options (Fig. 1): (1) restricting searches to cytoplasmic tails and loops of transmembrane proteins; (2) masking out structurally ordered regions (as predicted by IUPred [21] with a conservative threshold of 0.2) and/or relatively under-conserved residues [12]; (3) masking out domains, transmembrane and/or extracellular regions as annotated by UniProt [22]. Any combination of these options is permitted; users could, for example, restrict searches to cytoplasmic tails and loops of transmembrane proteins *and* mask out regions of predicted order, under-conserved residues and regions annotated as domains in UniProt.

3.3 Submitting Jobs

Once options have been chosen, clicking "Submit" will enter the job in the run queue. Run times will vary according to input data size and complexity, masking options and the current load of the server; the server has a maximum run time of 4 hours, after which jobs will be terminated. (For larger searches, users are encouraged to download and install a local version of SLiMSearch.) Each job is allocated a unique, randomly determined identifier. Users can either wait for their jobs to run, or bookmark the page and return to it later. Previously run job IDs can also be entered into a box on the SLiMSearch homepage to retrieve the run status and/or results.

3.4 Output

Once a job has run, the SLiMSearch results pages will open (Fig. 2). The main results page consists of a table of motif occurrences for each motif along with statistics for each occurrence including conservation (RLC) and disorder (IUPred). All fields can be sorted by clicking column headings and direct links to UniProt entries for each sequence are provided. The second primary results page consists of a summary table, which provides summary statistics for each motif. These include numbers of occurrences and SLiMChance assessments of over- or under-representation versus random expectation. Explanations of each field can be found in the SLiMSearch manual, which is available from the website. All the raw results files can also be downloaded for further analysis. When a user-defined dataset has been searched, these raw data files include the UniProt download. A key feature of SLiMSearch when analyzing user-defined datasets is the adjustment of the SLiMChance over- and under-representation statistics for evolutionary relatedness; for example, the probability of observing the Dynein Light Chain ligand "[KR].TQT" [28] in its annotated ELM proteins [2] *by chance* increases by eight orders of magnitude from 5.2e-18 to 4.2e-10 when the effective dataset size is reduced from 7 to 4 due to evolutionary relationships (Fig. 2). Whilst, in this example, the motif is still highly significant (the search dataset was defined based on the presence of the motif), in other cases this could be the difference between non-significance and apparent significance. Due to the size of the datasets, SLiMChance correction for evolutionary relationships is not available for human proteome searches.



SLiMSearch

Results

Motif Hits

Switch table view (Motifs|Summary)
Remove motifs with IUP less than 0.3|0.5|reset
Click on headers to sort

Motif Statistics

Click to switch motif.

Pattern	N_Occ
[KR].TQT	7

Viewing none

Seq	Desc	Pos	Len	RLC ↑	IUP	Pattern	Match
DC1I1	Cytoplasmic dynein 1 intermediate chain 1	151	628	2.03 view	0.563	[KR].TQT	KETQT
DC1I2	Cytoplasmic dynein 1 intermediate chain 2	158	638	1.95 view	0.605	[KR].TQT	KETQT
DYIN	Cytoplasmic dynein 1 intermediate chain	130	663	1.61 view	0.607	[KR].TQT	KQTQT
SWA	Protein swallow	283	537	1.52 view	0.474	[KR].TQT	KATQT
SWA	Protein swallow	291	548	1.4 view	0.532	[KR].TQT	KATQT
ZMY11	Zinc finger MYND domain-containing protein 11	413	562	0.616 view	0.477	[KR].TQT	RXTQT
B2L11	Bcl-2-like protein 11	112	198	0.518 view	0.624	[KR].TQT	KSTQT

Motif Hits

Pattern	IC	SeqNum	N_Occ	E_Occ	p_Occ	pUnd_Occ	N_Seq	E_Seq	p_Seq	pUnd_Seq	N_UPC	E_UPC	p_UPC	pUnd_UPC	Cons_mean	IUP_mean
[KR].TQT	3.77	7	7	0.024	8.60e-16	1.00	7	0.024	5.21e-18	1.00	4	0.018	4.19e-10	1.00	1.38	0.555

Raw Data

Fig. 2. SLiMSearch results pages. The main results page consists of a table of motif occurrences for each motif (top panel) along with statistics for each occurrence including conservation (RLC) and disorder (IUPred). All fields can be sorted by clicking column headings. Clicking sequence names will open the corresponding UniProt entry, while clicking “View” generates a visual representation of the motif. Clicking on different motifs in the smaller table on the left switches the motif being viewed. A summary table can also be viewed (bottom panel), which provides summary statistics for each motif. These statistics include SLiMChance assessments of over- or under-representation versus random expectation. Explanations of each field can be found in the SLiMSearch manual, which is available from the website. All the raw results files can also be accessed via the “Raw Data” link.

Individual motif occurrences can also be visualized for contextual information (Fig. 3). The multiple sequence alignment used for evolutionary conservation calculations is shown, with the relative conservation and IUPred disorder scores plotted below. Regions predicted to be ordered (below the disorder threshold of 0.2) are shaded, indicating areas that were (or would be) masked with disorder masking. In addition to these data, additional annotation from key SLiM and Protein databases is added. Annotated and unannotated Regular Expression matches to SLiMs from the Eukaryotic Linear Motif (ELM) database [2] are displayed above the alignment; sequence features from UniProt [22], including annotated domains and known mutations, are displayed between the alignment and RLC/Disorder plots. Users can hover the mouse over these features for additional information.

3.5 Getting Help

The SLiMSearch webserver is supported by an extensive help section, including a quickstart guide and walkthrough with screenshots. Example input files are provided. Fully interactive example output (corresponding to running the example Dynein Light Chain ligand input with default parameters) is clearly linked from the help pages. Additional details of the algorithms and options can be found in the SLiMSearch manual, which is also clearly linked from the help pages.

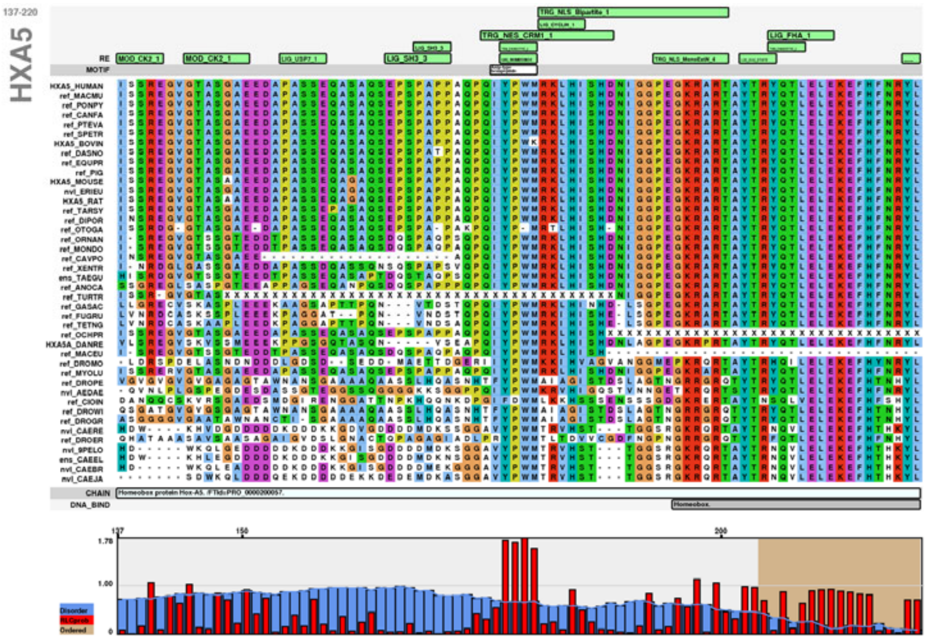


Fig. 3. Visualization of LIG_HOMEOBOX in HXA5 containing a multiple alignment of the orthologs of HXA5, drawn using Clustal coloring scheme, surrounded by relevant annotation. The bottom section contains a graph of relative conservation (in red) and IUPred disorder (in blue), with regions below the disorder threshold of 0.2 shaded (in brown). Above this section UniProt features are plotted, for example, in the case of HXA5 the right most region contains a DNA-binding Homeobox domain. Above the alignment, the motif row specifies regions containing a known functional motif (in white) and the RE row species regions matching the regular expression of a known motif (in green).

3.6 Server Limits

The server is currently limited to jobs with a run time of fewer than 4 hours. Motifs must have at least two non-wildcard positions defined and individual motif occurrence data is restricted to motifs with no more than 2000 occurrences in the search dataset. Custom UniProt datasets can have no more than 100 proteins. For larger analyses, users must install a local copy of the SLiMSearch software.

4 Example Analysis: HOX Ligand Motif

Homeobox (HOX) genes are a family of transcription factors controlling organization of segmental identity during embryo development [29] and recognized by a 60 residue DNA binding domain known as a Homeodomain [30]. HOX proteins recruit another Homeobox-containing transcription factor, PBX, via a conserved [FY][DEP]WM motif ("LIG_HOMEOBOX" [2]), binding a hydrophobic pocket created upon association of PBX to DNA [31]. Alone, the Homeodomain has weak specificity and affinity binding to the short DNA sequence TNAT, however following the formation of a heterodimer complex with TGAT binding PBX, bi-partite recognition increases specificity and allows HOX to specifically target developmental genes for expression.

A survey of the human proteome for [FY][DEP]WM PBX-binding motifs was completed to illustrate the effect of masking of globular regions and under conserved residues on the ability of a motif discovery tool to return functional motifs. Without any masking, SLiMSearch returned 53 motifs in 53 proteins, including the 16 annotated functional instances from the ELM database [2] (Supplementary Table 1). Of the 53 human occurrences, however, 30 were no longer returned following masking (IUPred masking cut-off 0.2, relative conservation filtering, domain masking and removal of extracellular and transmembrane regions). Of these 30, only 3 were known to be functional. The 23 remaining instances are all members of the Homeobox family; 13 of these contain a known annotated PBX-binding motif; given the homology of the remaining non-ELM containing proteins to the proteins containing function motifs, it is likely that all 23 instances are functional. The HXA5 occurrence, for example, shows a clear conservation signal characteristic of a functional motif despite not being annotated in ELM (Fig. 3).

5 Future Work

In addition to evolutionary conservation and structural disorder, successful identification of novel functional motifs in proteins can benefit from keyword or GO term enrichment [6, 32]. We are currently working on the incorporation of GO term enrichment into SLiMSearch analyses for future releases of the webserver. The current server is also limited to the human proteome only. In future we will expand this to include other organisms. Initially, these will be taken from the EnsEMBL database of eukaryotic genomes [33] and then expanded to other taxonomic groups [34]. We welcome suggestions from users, however, and will work with specific interest groups to add proteomes from appropriate species to the webserver where possible.

6 Conclusion

Discovering and annotating novel occurrences of Short Linear Motifs is an important ongoing task in biology, which often involves motif searches combined with additional evolutionary analyses (*e.g.* [32, 35]). The SLiMSearch webserver provides the biological community with an important advance in this arena, allowing evolutionary and structural context to be automatically incorporated into motif searches and visualized in user-friendly output. The flexibility of input, allowing known or novel motifs and user-defined protein datasets, combined with the statistical framework of SLiMChance for assessing motif abundance, makes SLiMSearch a powerful tool that should ease future discoveries of functional SLiM occurrences. In addition to the webserver implementation, SLiMSearch is available as standalone open source Python code under a GNU license, making it accessible to analyses of experimental biologists and bioinformatics specialists alike.

The SLiMSearch server is available at: http://bioware.ucd.ie/slimsearch.html. Supplementary Table 1 can be viewed at :http://bioware.ucd.ie/~compass/Server_pages/help/slimsearch/slimsearch_s1.pdf

Acknowledgments. This work was supported by Science Foundation Ireland, the University of Southampton and a European Molecular Biology Laboratory, EMBL Interdisciplinary Postdoc (EIPOD) fellowship (to N.E.D.).

References

1. Diella, F., Haslam, N., Chica, C., Budd, A., Michael, S., Brown, N.P., Trave, G., Gibson, T.J.: Understanding eukaryotic linear motifs and their role in cell signaling and regulation. Front Biosci. 13, 6580–6603 (2008)
2. Gould, C.M., Diella, F., Via, A., Puntervoll, P., Gemund, C., Chabanis-Davidson, S., Michael, S., Sayadi, A., Bryne, J.C., Chica, C., Seiler, M., Davey, N.E., Haslam, N., Weatheritt, R.J., Budd, A., Hughes, T., Pas, J., Rychlewski, L., Trave, G., Aasland, R., Helmer-Citterich, M., Linding, R., Gibson, T.J.: ELM: the status of the 2010 eukaryotic linear motif resource. Nucleic Acids Res. 38, D167–D180 (2010)
3. Kadaveru, K., Vyas, J., Schiller, M.R.: Viral infection and human disease–insights from minimotifs. Front Biosci. 13, 6455–6471 (2008)
4. Neduva, V., Russell, R.B.: Peptides mediating interaction networks: new leads at last. Curr. Opin. Biotechnol. 17, 465–471 (2006)
5. Rajasekaran, S., Balla, S., Gradie, P., Gryk, M.R., Kadaveru, K., Kundeti, V., Maciejewski, M.W., Mi, T., Rubino, N., Vyas, J., Schiller, M.R.: Minimotif miner 2nd release: a database and web system for motif search. Nucleic Acids Res. 37, D185–D190 (2009)
6. Ramu, C.: 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 (2003)
7. de Castro, E., Sigrist, C.J., Gattiker, A., Bulliard, V., Langendijk-Genevaux, P.S., Gasteiger, E., Bairoch, A., Hulo, N.: ScanProsite: detection of PROSITE signature matches and ProRule-associated functional and structural residues in proteins. Nucleic Acids Res. 34, W362–W365 (2006)
8. Gutman, R., Berezin, C., Wollman, R., Rosenberg, Y., Ben-Tal, N.: QuasiMotiFinder: protein annotation by searching for evolutionarily conserved motif-like patterns. Nucleic Acids Res. 33, W255–W261 (2005)
9. Bateman, A., Coin, L., Durbin, R., Finn, R.D., Hollich, V., Griffiths-Jones, S., Khanna, A., Marshall, M., Moxon, S., Sonnhammer, E.L., Studholme, D.J., Yeats, C., Eddy, S.R.: The Pfam protein families database. Nucleic Acids Res. 32, D138–D141 (2004)
10. Sigrist, C.J., Cerutti, L., de Castro, E., Langendijk-Genevaux, P.S., Bulliard, V., Bairoch, A., Hulo, N.: PROSITE, a protein domain database for functional characterization and annotation. Nucleic Acids Res. 38, D161–D166 (2010)
11. Chica, C., Labarga, A., Gould, C.M., Lopez, R., Gibson, T.J.: A tree-based conservation scoring method for short linear motifs in multiple alignments of protein sequences. BMC Bioinformatics 9, 229 (2008)
12. Davey, N.E., Shields, D.C., Edwards, R.J.: Masking residues using context-specific evolutionary conservation significantly improves short linear motif discovery. Bioinformatics 25, 443–450 (2009)

13. Via, A., Gould, C.M., Gemund, C., Gibson, T.J., Helmer-Citterich, M.: A structure filter for the Eukaryotic Linear Motif Resource. BMC Bioinformatics 10, 351 (2009)
14. Jonassen, I., Collins, J.F., Higgins, D.G.: Finding flexible patterns in unaligned protein sequences. Protein Sci. 4, 1587–1595 (1995)
15. Bailey, T.L., Boden, M., Buske, F.A., Frith, M., Grant, C.E., Clementi, L., Ren, J., Li, W.W., Noble, W.S.: MEME SUITE: tools for motif discovery and searching. Nucleic Acids Res. 37, W202–W208 (2009)
16. Neduva, V., Russell, R.B.: DILIMOT: discovery of linear motifs in proteins. Nucleic Acids Res. 34, W350–W355 (2006)
17. Davey, N.E., Shields, D.C., Edwards, R.J.: SLiMDisc: short, linear motif discovery, correcting for common evolutionary descent. Nucleic Acids Res. 34, 3546–3554 (2006)
18. Edwards, R.J., Davey, N.E., Shields, D.C.: SLiMFinder: A Probabilistic Method for Identifying Over-Represented, Convergently Evolved, Short Linear Motifs in Proteins. PLoS ONE 2, e967 (2007)
19. Edwards, R.J., Davey, N.E., Shields, D.C.: CompariMotif: quick and easy comparisons of sequence motifs. Bioinformatics 24, 1307–1309 (2008)
20. Davey, N.E., Haslam, N.J., Shields, D.C., Edwards, R.J.: SLiMFinder: a web server to find novel, significantly over-represented, short protein motifs. Nucleic Acids Res. (2010)
21. Dosztanyi, Z., Csizmok, V., Tompa, P., Simon, I.: IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. Bioinformatics 21, 3433–3434 (2005)
22. Bairoch, A., Apweiler, R., Wu, C.H., Barker, W.C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., Martin, M.J., Natale, D.A., O'Donovan, C., Redaschi, N., Yeh, L.S.: The Universal Protein Resource (UniProt). Nucleic Acids Res. 33, D154–D159 (2005)
23. Davey, N.E., Edwards, R.J., Shields, D.C.: The SLiMDisc server: short, linear motif discovery in proteins. Nucleic Acids Res. 35, W455–W459 (2007)
24. Russell, R.B., Gibson, T.J.: A careful disorderliness in the proteome: sites for interaction and targets for future therapies. FEBS Lett. 582, 1271–1275 (2008)
25. Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J.: Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25, 3389–3402 (1997)
26. Neuwald, A.F., Green, P.: Detecting patterns in protein sequences. J. Mol. Biol. 239, 698–712 (1994)
27. Seiler, M., Mehrle, A., Poustka, A., Wiemann, S.: The 3of5 web application for complex and comprehensive pattern matching in protein sequences. BMC Bioinformatics 7, 144 (2006)
28. Lo, K.W., Naisbitt, S., Fan, J.S., Sheng, M., Zhang, M.: The 8-kDa dynein light chain binds to its targets via a conserved (K/R)XTQT motif. J. Biol. Chem. 276, 14059–14066 (2001)
29. Wellik, D.M.: Hox genes and vertebrate axial pattern. Curr. Top Dev. Biol. 88, 257–278 (2009)
30. Gehring, W.J., Affolter, M., Burglin, T.: Homeodomain proteins. Annu. Rev. Biochem. 63, 487–526 (1994)
31. Sprules, T., Green, N., Featherstone, M., Gehring, K.: Lock and key binding of the HOX YPWM peptide to the PBX homeodomain. J. Biol. Chem. 278, 1053–1058 (2003)
32. Michael, S., Trave, G., Ramu, C., Chica, C., Gibson, T.J.: Discovery of candidate KEN-box motifs using cell cycle keyword enrichment combined with native disorder prediction and motif conservation. Bioinformatics 24, 453–457 (2008)

33. Hubbard, T.J., Aken, B.L., Ayling, S., Ballester, B., Beal, K., Bragin, E., Brent, S., Chen, Y., Clapham, P., Clarke, L., Coates, G., Fairley, S., Fitzgerald, S., Fernandez-Banet, J., Gordon, L., Graf, S., Haider, S., Hammond, M., Holland, R., Howe, K., Jenkinson, A., Johnson, N., Kahari, A., Keefe, D., Keenan, S., Kinsella, R., Kokocinski, F., Kulesha, E., Lawson, D., Longden, I., Megy, K., Meidl, P., Overduin, B., Parker, A., Pritchard, B., Rios, D., Schuster, M., Slater, G., Smedley, D., Spooner, W., Spudich, G., Trevanion, S., Vilella, A., Vogel, J., White, S., Wilder, S., Zadissa, A., Birney, E., Cunningham, F., Curwen, V., Durbin, R., Fernandez-Suarez, X.M., Herrero, J., Kasprzyk, A., Proctor, G., Smith, J., Searle, S., Flicek, P.: Ensembl 2009. Nucleic Acids Res. 37, D690–D697 (2009)
34. Kersey, P.J., Lawson, D., Birney, E., Derwent, P.S., Haimel, M., Herrero, J., Keenan, S., Kerhornou, A., Koscielny, G., Kahari, A., Kinsella, R.J., Kulesha, E., Maheswari, U., Megy, K., Nuhn, M., Proctor, G., Staines, D., Valentin, F., Vilella, A.J., Yates, A.: Ensembl Genomes: extending Ensembl across the taxonomic space. Nucleic Acids Res. 38, D563–D569 (2010)
35. Delpire, E., Gagnon, K.B.: Genome-wide analysis of SPAK/OSR1 binding motifs. Physiol Genomics 28, 223–231 (2007)