Nucleic Acids Research, 2007, Vol. 35, Web Server issue ***W455–W459***
doi:10.1093/nar/gkm400

The SLiMDisc server: short, linear motif discovery in proteins

Norman E. Davey, Richard J. Edwards and Denis C. Shields*

UCD Conway Institute of Biomolecular and Biomedical Research, University College Dublin, Belfield, Dublin 4, Ireland

Received January 29, 2007; Revised April 30, 2007; Accepted May 1, 2007

ABSTRACT

Short, linear motifs (SLiMs) play a critical role in many biological processes, particularly in protein–protein interactions. Overrepresentation of convergent occurrences of motifs in proteins with a common attribute (such as similar subcellular location or a shared interaction partner) provides a feasible means to discover novel occurrences computationally. The SLiMDisc (Short, Linear Motif Discovery) web server corrects for common ancestry in describing shared motifs, concentrating on the convergently evolved motifs. The server returns a listing of the most interesting motifs found within unmasked regions, ranked according to an information content-based scoring scheme. It allows interactive input masking, according to various criteria. Scoring allows for evolutionary relationships in the data sets through treatment of BLAST local alignments. Alongside this ranked list, visualizations of the results improve understanding of the context of suggested motifs, helping to identify true motifs of interest. These visualizations include alignments of motif occurrences, alignments of motifs and their homologues and a visual schematic of the top-ranked motifs. Additional options for filtering and/or re-ranking motifs further permit the user to focus on motifs with desired attributes. Returned motifs can also be compared with known SLiMs from the literature. SLiMDisc is available at: http://bioware.ucd.ie/~slimdisc/.

INTRODUCTION

Short, linear motifs (SLiMs) play an essential role in the basic functions of many proteins and identifying these motifs is of great interest in expanding our understanding of biological interaction networks (1). They facilitate many fundamental biological tasks, such as subcellular targeting (e.g. The DxxLL Endosome-Lysosome-Basolateral sorting signal motif), post-translational modification (e.g. The NxC N-Linked glycosylation site motif) and protein binding [e.g. The (KR)xTQT Dynein Light Chain binding motif](2). However, our current knowledge of the area is sparse with less than a hundred classes of SLiMs known in eukaryotes. Estimates have been made that suggest hundreds more are still undiscovered, making the area of SLiM discovery worthy of substantial investment of effort (3).

A powerful way to discover novel SLiMs comes from the fact that they are evolutionarily plastic, making them amenable to convergent evolution. SLiMs of this type can be found by searching for motifs which are shared between proteins with a common attribute (such as biological function, subcellular location or a common interaction partner) and have no evidence of a shared ancestry. This approach has been implemented for motif discovery in the LMD/DILIMOT (4,5) and Short, Linear Motif Discovery (SLiMDisc) methods (6), which incorporate both masking and evolutionary filtering to return overrepresented motifs. The principal underlying difference between the two methods is that DILIMOT masks all but one arbitrarily selected homologous protein prior to motif discovery, whereas SLiMDisc searches for motifs in all proteins and then weights results according to the evolutionarily relationships of the proteins containing the motif.

This article describes a web-based application of the SLiMDisc method, adding interactive masking of protein features, useful data visualizations and additional ranking and filtering options compared with the standalone program (6). While interpretation of motifs requires first a ranked scoring of the most significant motifs, the context of the motif is of great importance in evaluating how likely it is to be of interest to investigate further. SLiMDisc permits rapid visualization of aspects of motif context, such as possible weak, uncorrected evolutionary relationships between proteins sharing the motif, the level of disorder of the protein

*To whom correspondence should be addressed. Tel: +353-1-7165344; Fax: +353-1-7166701; Email: denis.shields@ucd.ie

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

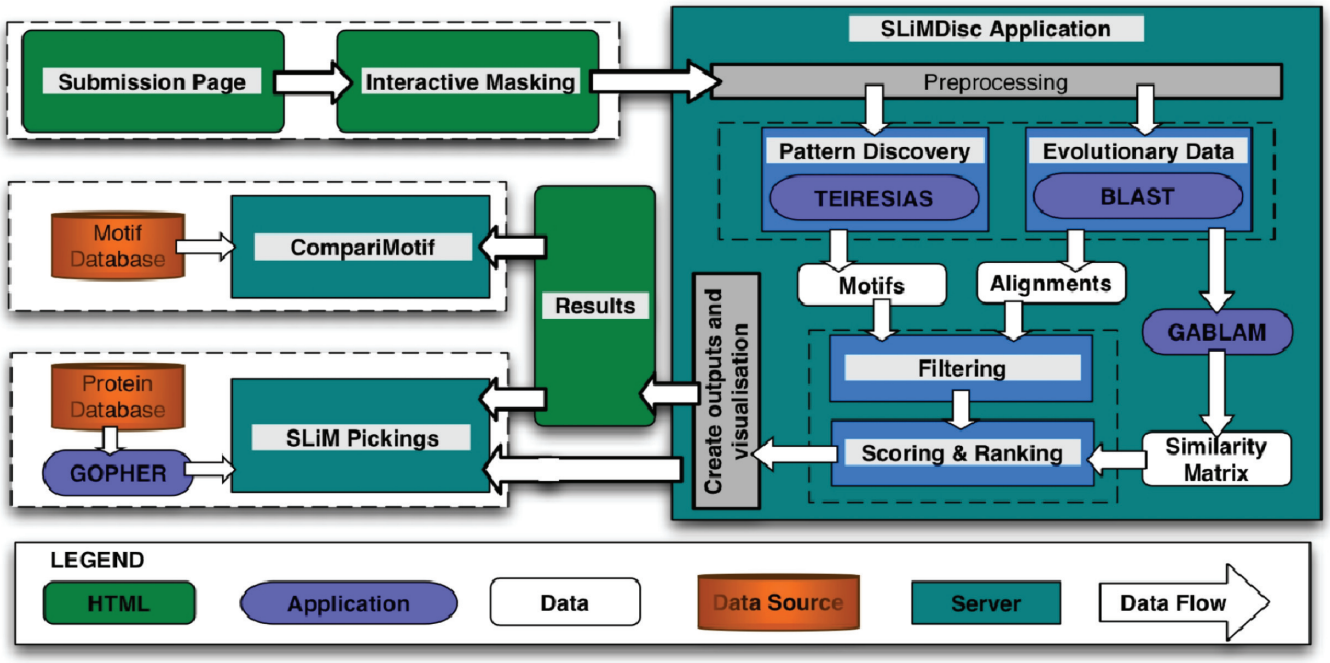

Figure 1. Schema of the SLiMDisc server. The SLiMDisc server returns putatively functional motifs contained in the input data set. CompariMotif and SLiMPickings post-process the output of the SLiMDisc server adding functionality such as conservation scoring, re-ranking and comparison against known motifs.

region (often motifs are enriched in disordered regions) and the evolutionary divergence of the motif in homologues.

Examples of SLiMDisc applied to interaction data sets have been discussed previously when it was shown to successfully recover known shared linear motifs from the HPRD-curated protein interaction database (6). Here we describe the SLIMDisc web server, which implements this stand-alone program, and provides in addition important visualization tools to aid interpretation of the results. This article illustrates the flexibility of the SLiMDisc server through an example application to a real biological data set.

METHOD OVERVIEW

Concept

SLiMDisc is a motif discovery method for finding putative functional motifs in a group of proteins with a common attribute (6). The strongest evidence for a functional motif is obtained when the same motif occurs in unrelated proteins, evolving by convergence. Searches for such motifs are often swamped by motifs that are shared in related proteins and identical by descent; SLiMDisc incorporates information about the evolutionary descent of each motif based on treatment of BLAST (7) local alignments to weight against this effect.

SLiMDisc overview

SLiMDisc outputs a ranked list of putatively interesting SLiMs from an input data set of proteins. A schema of the method is given in Figure 1. Any protein regions defined by users to be masked are removed from the data sets. Homology information for the data set is created using the GABLAM method (6). TEIRESIAS (8) is used to identify all motifs with or without ambiguity. Motifs are filtered according to a number of optional criteria, including the evolutionary relatedness of the proteins containing the motif, information content and surface probability. Finally, the motifs are ranked according to information content and support, normalized for evolutionary relationships (6).

Masking

Masking of regions unlikely to contain motifs can greatly improve SLiM discovery. It has been observed that functional motifs occur much less frequently inside globular regions, coiled coils and, by definition, in regions inaccessible for interaction (2,5). It is clear that removal of these regions will decrease the size of the search space, thus decreasing noise introduced by randomly occurring shared patterns in the data set and increasing the likelihood of discovering any true, functional SLiMs.

The SLiMDisc server allows a user to easily mask regions that, from prior knowledge of the data set, they believe to be unlikely to contain functional motifs. Regions such as transmembrane regions, protein domains and inaccessible residues can be masked since, for certain purposes, they are areas which may have a lower likelihood of containing motifs. Furthermore, the user may identify specific regions of the proteins in which to confine the search (e.g. the cytoplasmic regions of a set of proteins) or to confine the search based on prior knowledge of a region possibly containing a functional

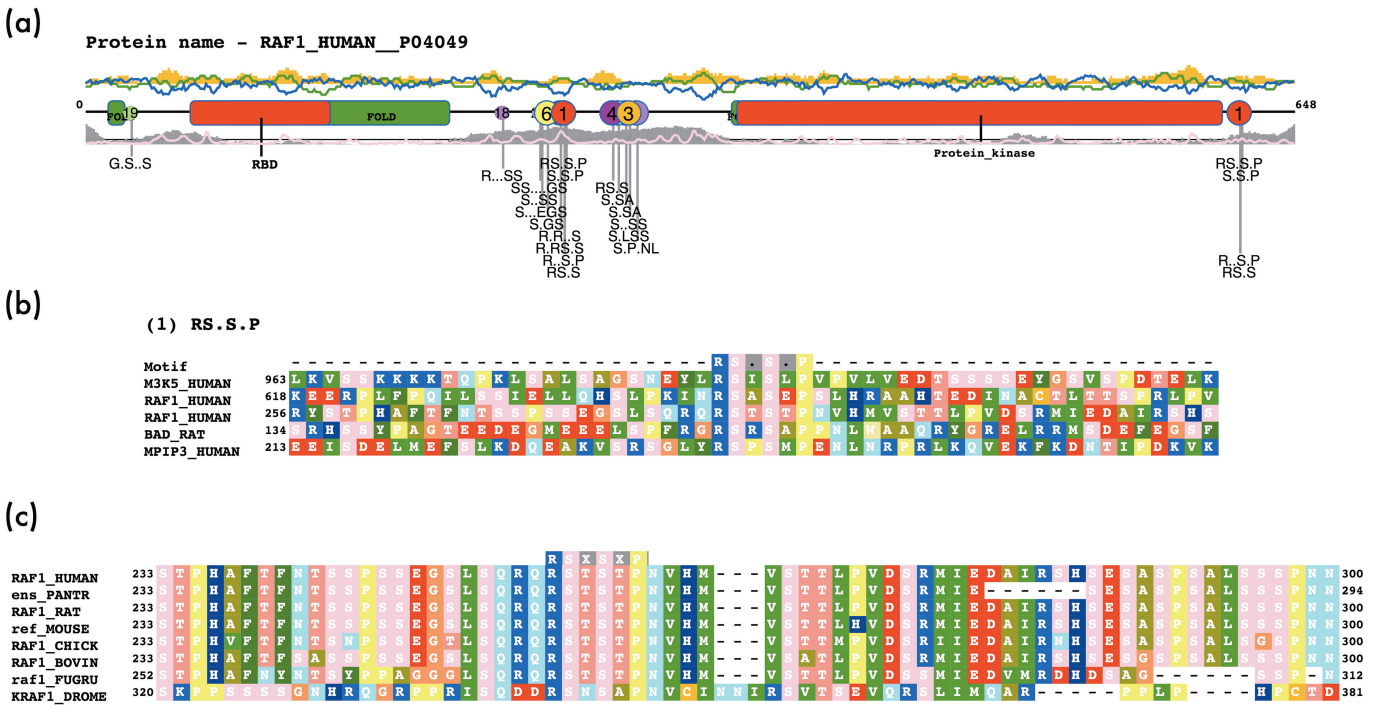

Figure 2. Visualizations created by the SLiMDisc web server for proteins from the ELM database (2) with annotated 14-3-3 binding motifs. **(a)** A motif position map shows the position of the top 20 returned motifs relative to the UNIPROT annotated features of the protein. Returned patterns are displayed below each motif marker, enabling the user to easily identify similar, overlapping motifs. The graph includes several charge-related scores, Eisenberg hydropathy and protein region disorder as predicted by IUPRED. The graph also shows features which have been exclusively masked according to the user's preferences (red: known domains; green: labelled 'FOLD', indicates other regions that IUPRED predicted to have a low disorder, i.e. are more likely to be folded). SA: Surface accessibility. **(b)** An ungapped alignment of the residues immediately surrounding the motif returned, showing the context of each occurrence. **(c)** An alignment (allowing gaps) of the orthologues of the RAF1 protein around the RSxSxP 14-3-3 binding motif. This visualization shows the conservation of the motif across several species.

motif in a given protein. This allows easy incorporation of prior knowledge without lengthy post-processing of results and also reduces search times.

SLiM scoring

Having established a large number of motifs, SLiMDisc ranks them according to their information content and frequency of occurrence, adjusted for evolutionary relationships (6). This scoring method has been shown to perform well in the rediscovery of known ELMs in protein–protein interaction data sets where a subset of the data set is known to contain a functional ELM. The support or number of occurrences of each motif is down-weighted according to the relatedness of the proteins containing the motif. Three methods are available for the calculation of the normalized support of a motif [Minimum Spanning Tree, Unique Homologous Segments and Unrelated Proteins normalization (described in previous work)], each of which has a varying level of strictness with regards to the treatment of relationships between proteins.

Output

SLiMDisc outputs a ranked list of putatively interesting motifs. In addition, the server produces several visualizations (in graphical PDF format) that allow users to gain knowledge of the context of the motif. These include: protein similarity trees using the GABLAM algorithm (6); ungapped alignments of the regions of proteins containing each motif centred on the motif; alignments (allowing gaps) of orthologues of the proteins, generated by MUSCLE (9); motif maps showing the position of the motifs relative to other features (e.g. domains, transmembrane regions) in the proteins, including plots of charge, surface probability, hydrophobicity and disorder [based on an 11-amino acid window size (Figure 2)]. All outputs are also produced as plain text for easy use with other programs.

Post-processing

Two additional servers for post-processing of the SLiMDisc results are available upon completion of a SLiMDisc analysis: SLiMPickings and CompariMotif (see Supplementary Materials for details). Submission to each of these servers is achieved by pressing a single button on the SLiMDisc results page, and a range of post-processing options is available if the user wishes to depart from the default settings.

The SLiMPickings server allows for easy re-ranking and post-filtering of data based on user-defined options. The server compiles data from SLiMDisc and analyses it to create various statistics such as SLiM conservation (conservation scores are calculated on orthologues returned from the GOPHER server), predicted region disorder using IUPRED (10) and several charge-related

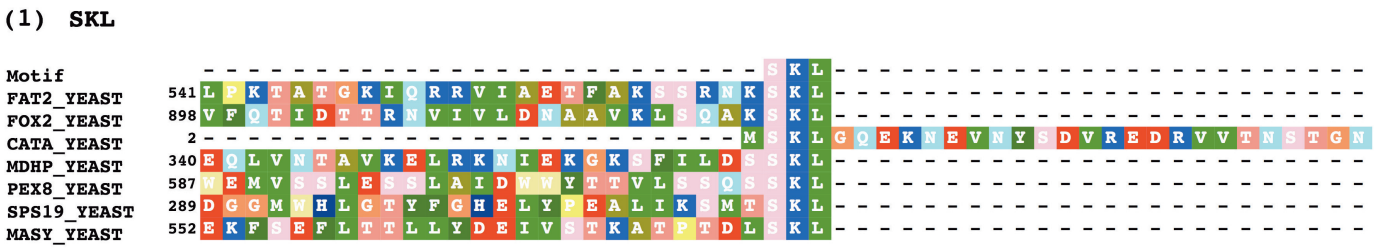

Figure 3. Alignment of proteins containing the SKL motif found in the 'peroxisomal matrix' Gene Ontology data set. The shared position of the motif near the protein termini improves our overall confidence in the true functionality of the motif.

statistics (such as absolute charge, net charge and charge balance) for a user-defined window size around each occurrence. The original SLiMDisc results can be re-ranked and/or filtered on any of the statistics produced by the server, including novel scoring schemes created by the user.

The CompariMotif server uses motif regular expressions to match putative motifs returned from the SLiMDisc analysis against libraries of known SLiMs. Libraries available include motif definitions drawn from the ELM database (2,version, 28 June 2005) and the MiniMotif database (11). The server allows flexibility in motif–motif comparison, permitting incomplete matches ('*x*' of the '*n*' positions in the motif must match), overlapping matches (neither motif is entirely contained within the other) and parent matches (one motif is longer and entirely contains the other motif). Ambiguity is also considered in the comparisons. Output is a sortable table of all pairs of matching motifs, along with their degree of similarity (based on information content) and their relationship to each other.

Server details

The SLiMDisc server is a Python-CGI server with a first-in first-out queuing system. The input is a data set of protein sequences which have been hypothesized to have a possible common motif due to a common attribute such as biological function, subcellular location or a common interaction. UNIPROT flat file (12) and FASTA formats are fully supported. Most other major formats are also permitted.

All the functionality of the standalone SLiMDisc method is available through the advanced user options in the web interface; however, these are set to the default and are therefore discretionary. The main advantages of the web-based SLiMDisc over the standalone version are the inclusion of interactive protein region masking, data visualizations, and post-processing ranking, filtering and motif comparison options.

Jobs take different lengths of time depending on data set size and evolutionary relationship between proteins. Certain data sets will not be processed if they are estimated to be overly computationally expensive. In such cases, the user will be notified and will be advised to use the local version of the method. Post-processing and visualization tools may be used by uploading files generated using a local version of the SLiMDisc application.

EXAMPLE OF USAGE

Peroxisomal targeting signal (PTS)

All peroxisomal proteins are synthesized in the cytoplasm before being transported to the peroxisome. The majority of peroxisomal proteins contain a PTS that binds a peroxin, Pex5, which mediates transport across the peroxisomal membrane to the peroxisomal matrix. We created a data set of 19 proteins by taking all Gene Product Associations for the cell component 'peroxisomal matrix' from the Gene Ontology database (13) and analysed it for any over-represented convergently evolving motifs. The top motif returned from the search was the motif SKL, occurring in eight of the proteins (with normalized MST support of 6.9). This motif is the archetypal, functional, peroxisomal targeting signal to which the peroxins bind before facilitating transport (14).

Ignoring the context of this motif, it does not stand out very strongly compared with several other overrepresented motifs, with slightly lower rankings. However, upon inspection of the visualizations for the data set, all occurrences of the motif can be seen to occur within five residues of either the C-terminus or the N-terminus (Figure 3). Visualization in this case allows the user to see the context of the motif greatly improving their confidence in the motif. The shared position of the motif at the C- or N- termini of seven proteins, from which only two have any evidence of common ancestry, is compelling support for true functionality.

The classic PTS motif is serine-lysine-leucine (14); however, each position allows the substitution of similar residues so the true consensus motif is *[SAPTC] [KRH][LMFI]*. In an attempt to improve upon the motif descriptor for the targeting signal returned from the data set, the interactive masking feature of the web server was used to restrict analysis to the 20 residues nearest the C- and N-termini of the proteins and the analysis re-run allowing ambiguous positions (with equivalence groupings: KR, DE, ILMV, AG, QN, FYW, ST). The top five ranked motifs returned all describe PTS variants of which the third ranked motif, S[KR][ILMV], is the best descriptor, seen in 12 of the proteins (with normalized MST support 10.8). The data set and result files including visualizations for both analyses are available at the SLiMDisc website: http://bioware.ucd.ie/~slimdisc/example/peroxisome.html.

DISCUSSION

We have created a web server for the SLiMDisc method for the discovery of putatively interesting motifs in protein data sets which haves some common attributes which may be explained by a shared motif. We introduced interactive masking (inclusive and exclusive) of annotated protein features on a protein by protein basis to decrease the search space and increase the likelihood of returning motifs of interest. We have also demonstrated how the use of visualization can be used to help distinguish biologically meaningful SLiMs from the noise of stochastically occurring background motifs.

SLiM discovery is a difficult task that has often been compared with finding a needle in a haystack. Their short length, high ambiguity and low-binding affinities makes them unfavourable for experimental or *in silico* discovery, and the lack of adequate training data sets makes it hard to develop robust discovery algorithms that are fully automated. Current methods are only likely to discover the most easily revealed motifs. At this time, SLiM discovery difficulties are best overcome by a combination of automated discovery followed by visual inspection of putatively functional motifs, and the supervised incorporation of as much expert knowledge as possible through data set masking and results filtering. Motif attributes, such as motif conservation in close orthologues or common motif position relative to other protein features, can all point to true functionality that is most easily described by visualization. The overall context of a motif allows the user to gain confidence in motifs and helps the user to separate true motifs from the background of randomly occurring motifs with high support. As more SLiMs are discovered and confirmed experimentally, it is hoped that our overall knowledge of the area will increase and allow more sophisticated—and more accurate—algorithms to be developed.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

This work was supported by Science Foundation Ireland. Funding to pay the Open Access publication charges for this article was provided by Science Foundation Ireland.

Conflict of interest statement. None declared.

REFERENCES

1. Neduva,V. and Russell,R.B. (2005) Linear motifs: evolutionary interaction switches. *FEBS Lett.*, **579**, 3342–3345.
2. Puntervoll,P., Linding,R., Gemund,C., Chabanis-Davidson,S., Mattingsdal,M., Cameron,S., Martin,D.M., Ausiello,G., Brannetti,B. *et al.* (2003) ELM server: A new resource for investigating short functional sites in modular eukaryotic proteins. *Nucleic Acids Res.*, **31**, 3625–3630.
3. Neduva,V. and Russell,R.B. (2006) Peptides mediating interaction networks: new leads at last. *Curr. Opin. Biotechnol.*, **17**, 465–471.
4. Neduva,V. and Russell,R.B. (2006) DILIMOT: discovery of linear motifs in proteins. *Nucleic Acids Res.*, **34**, W350–W355.
5. Neduva,V., Linding,R., Su-Angrand,I., Stark,A., de Masi,F., Gibson,T.J., Lewis,J., Serrano,L. and Russell,R.B. (2005) Systematic discovery of new recognition peptides mediating protein interaction networks. *PLoS Biol.*, **3**, e405.
6. Davey,N.E., Shields,D.C. and Edwards,R.J. (2006) SLiMDisc: short, linear motif discovery, correcting for common evolutionary descent. *Nucleic Acids Res.*, **34**, 3546–3554.
7. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
8. Rigoutsos,I. and Floratos,A. (1998) Combinatorial pattern discovery in biological sequences: The TEIRESIAS algorithm. *Bioinformatics*, **14**, 55–67.
9. Edgar,R.C. (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*, **5**, 113.
10. Dosztanyi,Z., Csizmok,V., Tompa,P. and Simon,I. (2005) IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. *Bioinformatics*, **21**, 3433–3434.
11. Balla,S., Thapar,V., Verma,S., Luong,T., Faghri,T., Huang,C.H., Rajasekaran,S., del Campo,J.J., Shinn,J.H. *et al.* (2006) Minimotif Miner: a tool for investigating protein function. *Nat. Methods*, **3**, 175–177.
12. Bairoch,A., Apweiler,R., Wu,C.H., Barker,W.C., Boeckmann,B., Ferro,S., Gasteiger,E., Huang,H., Lopez,R. *et al.* (2005) The Universal Protein Resource (UniProt). *Nucleic Acids Res.*, **33**, D154–D159.
13. Gene Ontology Consortium. (2006) The Gene Ontology (GO) project in 2006. *Nucleic Acids Res.*, **34**, D322–D326.
14. Gould,S.J., Keller,G.A., Hosken,N., Wilkinson,J. and Subramani,S. (1989) A conserved tripeptide sorts proteins to peroxisomes. *J. Cell. Biol.*, **108**, 1657–1664.