Nucleic Acids Research Advance Access published May 26, 2011

Nucleic Acids Research, 2011, 1–5
doi:10.1093/nar/gkr402

SLiMSearch 2.0: biological context for short linear motifs in proteins

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]UCD Complex and Adaptive Systems Laboratory, UCD Conway Institute, and School of Medicine and Medical Sciences, University College Dublin, Dublin 4, Ireland and [3]School of Biological Sciences, University of Southampton, Southampton, UK

Received February 12, 2011; Revised April 26, 2011; Accepted May 5, 2011

ABSTRACT

Short, linear motifs (SLiMs) play a critical role in many biological processes. The SLiMSearch 2.0 (Short, Linear Motif Search) web server allows researchers to identify occurrences of a user-defined SLiM in a proteome, using conservation and protein disorder context statistics to rank occurrences. User-friendly output and visualizations of motif context allow the user to quickly gain insight into the validity of a putatively functional motif occurrence. For each motif occurrence, overlapping UniProt features and annotated SLiMs are displayed. Visualization also includes annotated multiple sequence alignments surrounding each occurrence, showing conservation and protein disorder statistics in addition to known and predicted SLiMs, protein domains and known post-translational modifications. In addition, enrichment of Gene Ontology terms and protein interaction partners are provided as indicators of possible motif function. All web server results are available for download. Users can search motifs against the human proteome or a subset thereof defined by Uniprot accession numbers or GO term. The SLiMSearch server is available at: http://bioware.ucd.ie/slimsearch2.html.

INTRODUCTION

The purpose of the SLiMSearch (Short, Linear Motif Search) web server is to allow researchers to identify novel occurrences of user-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) through post-translational modification (including cleavage), subcellular localization and ligand binding (2). Once a SLiM has been defined, finding matches in a given set of protein sequences is a fairly trivial task. Several web-based methods to discover novel instances of known SLiMs are available, including ELM (2), MnM (3), SIRW (4) ScanProsite (5) and QuasiMotifFinder (6), which generally utilize databases of known motif patterns to search query protein sequences supplied by the user.

While finding matches is trivial, however, interpreting their biological significance is far from easy. Stochastic occurrences of small, degenerate motifs are 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 (5), for example, has an option to 'Exclude motifs with a high probability of occurrence', while QuasiMotifFinder (6) uses the background occurrence of motifs in PfamA families (7) to assess the significance of hits. These strategies work well for longer, family descriptor motifs [such as are found in the Prosite database (8) 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 (3,6,9,10), structural context (3,11) or even biological keywords (4) 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 (3) or Prosite (8). Those that permit users to define their own motifs, such as ScanProsite (5), 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 (12), MEME (13), Dilimot (14), SLiMDisc (15),

*To whom correspondence should be addressed. Tel: +353 1 716 5344; Email: denis.shields@ucd.ie
Correspondence may also be addressed to Richard J. Edwards. Tel: 02380594344; Fax: 023 8059 4459; Email: r.edwards@southampton.ac.uk

SLiMFinder (16) and FIRE-pro (17)]. Although SLiMFinder (16) 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. One approach is to compare candidate SLiMs to existing motif libraries to identify similarities to previously known motifs (18).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 (16), available as a web server (19). To aid interpretation of SLiMFinder results, we made a new tool available, SLiMSearch, which allows users to search protein data sets with user-defined motifs, including motif prediction output from SLiMFinder (20). SLiMSearch utilized 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 (10,21) and features the same SLiMChance algorithm for assessing statistical overrepresentation of SLiM occurrences, correcting for biases introduced by evolutionary relationships within the data (16). Like SLiMFinder, SLiMSearch was optimized for small protein data sets. In this article, we describe a complementary server, SLiMSearch 2.0, which is optimized for searches of a whole proteome.

SLiMSearch 2.0 replaces SLiMChance data set probabilities with individual likelihoods for each motif instance that permit the ranking of many motif occurrences and helps separate putative functional instances from the background of stochastic occurrences. A comprehensive study of the Eukaryotic Linear Motif (ELM) database by Fuxreiter *et al.* (22) found that SLiMs are more likely to be found in disordered regions, while Chica *et al.* (9) found that conserved motifs are more likely to be true positives. Our previous work with both discovery of new instances of known motifs and of novel motifs shows that, motifs in disordered regions and conserved motifs are typically (but not always) more likely to be true positives (10). Therefore, we encourage the use of an optional disorder filter and we present the results ranked according to conservation. Enrichment scores for motif counts are calculated (i) versus reversed/shuffled variants of the motif, (ii) for Gene Ontology (GO) terms (23) and (iii) for known BioGRID interactors of individual hub proteins (24). In addition to identifying individual occurrences of known motifs, therefore, SLiMSearch 2.0 can indicate possible functional significance for entirely novel motifs. Input, output and results visualizations are fully compatible with our existing SLiM analysis web servers, SLiMDisc (25), CompariMotif (18), SLiMFinder (19) and SLiMSearch 1.0 (20), providing a suite of integrated tools for analysing these biologically important sequence features.

THE SLIMSEARCH 2.0 ALGORITHM

SLiMSearch 2.0 performs a motif regular expression search against a proteome allowing restriction of considered sequences to set of proteins or a given GO term. Features include annotation of overlapping sequence annotation and calculation of global and local motif statistics and attributes.

Pre-formatted database

To speed up motif attribute calculations, pre-computed databases for each proteome are used. The current release has only Human UniProt release v1.37 (Aug 2010) (26); however, more model proteomes will be added as data is computed. Two pre-computed conservation scores are calculated for each protein in the proteome, a column-based tree-weighted conservation score (WCS) (9) and a relative local conservation (RLC) metric (10). Homologues for each sequence are identified using a BLAST search against a database of 70 complete EnsEMBL proteomes (Ensembl 59, October 2010, 69 Metazoan proteomes and *Saccharomyces cerevisiae*) (27) and orthologues are predicted using GOPHER (default options) (25). Predicted orthologues are aligned by MAFFT (28) and used to calculate conservation scoring metrics on a residue-by-residue basis. Disorder scores for each residue are calculated using IUPred (default options) (21).

Several features of interest are also preformatted for rapid querying: (i) Domain data from Pfam (29); (ii) structure data from PDB (30); (iii) experimentally validated motifs from the ELM database (2); and (iv) SNP and modification data from UniProt annotation (26).

Scoring

The IUPred disorder score, IUP, of the motif is calculated as the mean disorder score across the defined (non-wildcard) residues. The WCS of a motif is calculated similarly. SLiMSearch 2.0 extends the RLC score to return a probability. Based on the assumption, consistent with empirical observation, that the RLC scores for a residue are normally distributed (10), the RLC of a residue is converted into a probability, P(RLC), using the Gaussian Cumulative Distribution Function (CDF). The relative conservation probability of a motif, P, the probability of each residue of a motif having its given RLC or higher can be calculate as the product of the P(RLC) for each residue within the motif. A significance value, P(cons), representing the probability of a given motif having that P-value or lower by chance, can then be calculated for the motifs P-value using the CDF of the uniform product distribution [Equation (1)]. Thus, the P(cons) statistic provides a useful measure of how likely it is that this motif will have the observed degree of local conservation (or higher) by chance. Note that it does not provide any indication of the probability of the motif itself, which is best inferred from the enrichment values.

$$P(\text{cons}) = \frac{(-1)^n(-\ln(P))^{-n}\ln(P)^n\Gamma(n, -\ln(P))}{(n-1)!} \quad (1)$$

P(cons) is the probability of a given motif having that P-value or higher by chance, calculated as the CDF of the uniform product distribution, i.e. the distribution of the product of n uniform distributions, where n is the number of non-wildcard positions in the motif, P is the relative conservation probability of a motif and Γ is the incomplete gamma function.

Enrichment scores

Enrichment scores for motif counts are calculated for the input motif against the reverse of the motif and a randomly shuffled variant of the motif. The score is a simple quotient, where the input motif count is the divisor and the shuffled or reverse motif count is the dividend. Enrichment scores for each GO term and BioGRID interaction hub protein are calculated versus the expectation provided by the whole proteome, i.e. the number of motif occurrences in proteins with that GO term/interaction partner divided by the expected number of proteins, which is the total number of proteins with the motif multiplied by the proportion of the proteome with that GO term/interaction partner. Enrichment significance is calculated using the Fisher's exact test. Counts are normalized for independence by clustering highly similar proteins based on UniRef50 groups.

THE SLIMSEARCH 2.0 WEBSERVER

The SLiMSearch 2.0 server is available at http://bioware.ucd.ie/slimsearch2.html. The website is free and open to all and there is no login requirement. The purpose of the web server is to allow researchers to identify novel occurrences of user-defined SLiMs in a set of sequences. A rapid pattern matching search is first performed to identify all occurrences of the motif in the proteome (or a defined subset). Pre-formatted databases are then used to rapidly extract scores and sequence features for each occurrence before enrichment scores are calculated. Interactive output and visualizations permit easy exploration of returned occurrences of the motif and their sequence context. These features of the web server are described in more detail in the following sections.

Input

A motif to be compared against the search is the sole compulsory input. The motif should be expressed as a regular expression using single letter amino acid codes (e.g. R.LF or RxLF but not Arg–x–Leu–Phe). The format allows for ambiguity (i.e. positions that can be any residue from a set of residues, e.g. [ILV] meaning any aliphatic residue), flexibility (e.g. '.{1,3}', meaning a wildcard position between 1 and 3 residues in length), termini definition (where ^ is the N-terminus and $ is the C-terminus) and conditional motifs (e.g. (motif1)|(motif2) meaning motif1 or motif2). Two optional filtering options are also available, restricting the protein search space to a subset of a proteome: by GO term (in the format GO:0005868) to restrict the search to a particular ontology and similarly, to a set of proteins by UniProt accessions. For clarity, example inputs are available above each entry box on the input page of the web server.

Submitting jobs

Once input has been determined, clicking 'Submit job' will enter the run queue. Run times will vary according to input data size, motif complexity and the current load of the server but are generally in the order of a few seconds. Users can either wait for their jobs to run or bookmark the page and return to it later, although jobs are deleted after 21 days. The web server can also be run directly using a URL containing the motif to be searched and (optionally) a list of UniProt IDs.

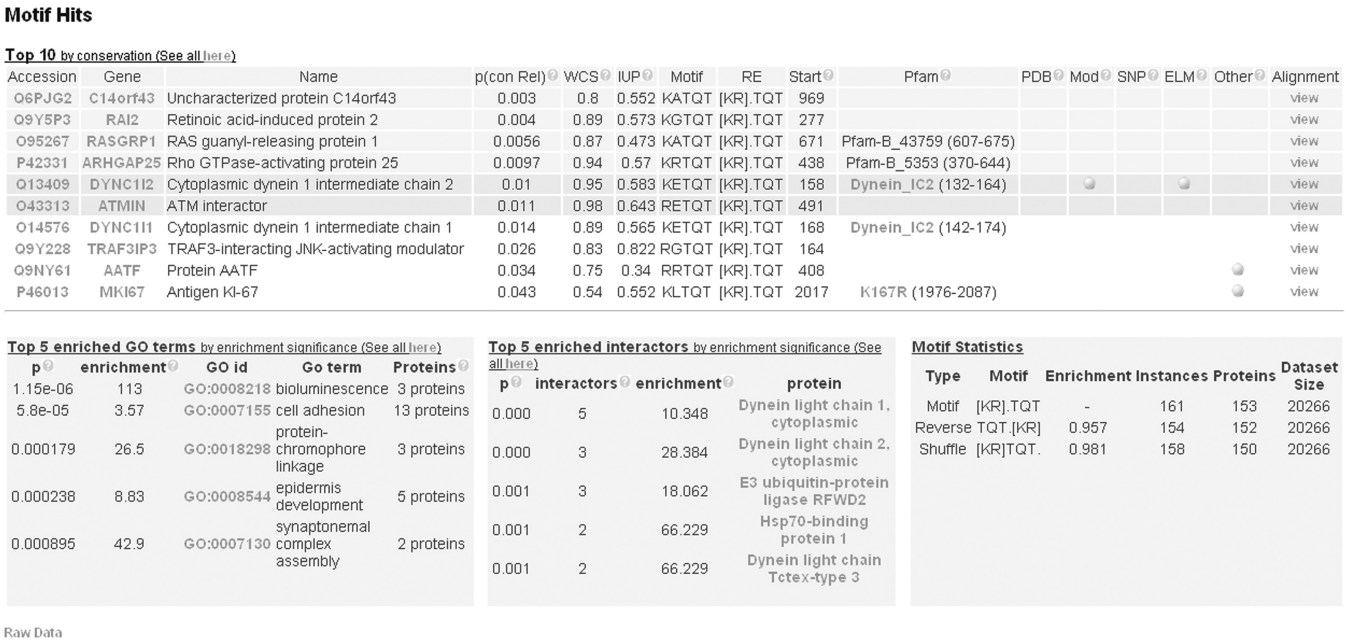
Output

The main output is a table of motif instances annotated with attributes including: (i) conservation and disorder statistics; (ii) overlapping feature, such as Pfam domains, PDB structures, SNPs and modifications; and (iii) overlapping experimentally validated motifs (Figure 1). In addition, alignments of 100 amino acid regions overlapping each motif occurrence can be visualized. Discovered motifs are not filtered, therefore all instances are returned. By default, motifs are ranked based on P(cons). Several additional tables are also returned: GO terms which are enriched for the motif; hub proteins where the interactors are enriched for the motif and motif count statistics. All results are returned as tab-delimited files and in a more visually appealing html format. Initially, an overview of the most interesting instances and enrichments are returned. More detailed data are available and can be sorted by each attribute. Instance data can be also filtered based on IUPred mean disorder score, IUP.

Users need to consider two separate lines of evidence when assessing the significance or otherwise of the findings presented. First, the motif enrichment over the reversed and shuffled sequences gives an indication to what extent the motifs that are provided occur by chance. If a motif occurs 40 times and the reverse occurs 20 times, this means that we expect that about half of the observed instances are false positives (assuming no negative selection on randomly occurring motifs). The user can then scroll down the list of occurrences, and investigate the conservation values, to form a judgement regarding which motifs are most likely to be true positives. Assuming a typical mammalian motif, it would be expected in this case that the 20 least conserved motifs are most likely to be false positives and the 20 most conserved are most likely to be true positives. In many cases, the enrichment may be relatively modest; P(cons) only provides guidance, rather than proof, regarding the likelihood that a given motif occurrence is a true positive.

Example analysis

The web server incorporates a full example for searching the human proteome with the manually curated, experimentally validated, Dynein Light Chain binding motif ([KR].TQT; ELM entry LIG_Dynein_DLC8_1 (2)). A full walkthrough for this data set is provided in the help pages and fully interactive example output is also



Motif Hits

Top 10 by conservation (See all here)

Accession	Gene	Name	p(con Rel)	WCS	IUP	Motif	RE	Start	Pfam	PDB	Mod	SNP	ELM	Other	Alignment
O6PJG2	C14orf43	Uncharacterized protein C14orf43	0.003	0.8	0.552	KATQT	[KR].TQT	969							view
O9Y5P3	RAI2	Retinoic acid-induced protein 2	0.004	0.89	0.573	KGTQT	[KR].TQT	277							view
O95267	RASGRP1	RAS guanyl-releasing protein 1	0.0056	0.87	0.473	KATQT	[KR].TQT	671	Pfam-B_43759 (607-675)						view
P42331	ARHGAP25	Rho GTPase-activating protein 25	0.0097	0.94	0.57	KRTQT	[KR].TQT	438	Pfam-B_5353 (370-644)						view
Q13409	DYNC1I2	Cytoplasmic dynein 1 intermediate chain 2	0.01	0.95	0.583	KETQT	[KR].TQT	158	Dynein_IC2 (132-164)						view
O43313	ATMIN	ATM interactor	0.011	0.98	0.643	RETQT	[KR].TQT	491							view
O14576	DYNC1I1	Cytoplasmic dynein 1 intermediate chain 1	0.014	0.89	0.565	KETQT	[KR].TQT	168	Dynein_IC2 (142-174)						view
O9Y228	TRAF3IP3	TRAF3-interacting JNK-activating modulator	0.026	0.83	0.822	RGTQT	[KR].TQT	164							view
O9NY61	AATF	Protein AATF	0.034	0.75	0.34	RRTQT	[KR].TQT	408							view
P46013	MKI67	Antigen KI-67	0.043	0.54	0.552	KLTQT	[KR].TQT	2017	K167R (1976-2087)						view

Top 5 enriched GO terms by enrichment significance (See all here)

p	enrichment	GO id	Go term	Proteins
1.15e-06	113	GO:0008218	bioluminescence	3 proteins
5.8e-05	3.57	GO:0007155	cell adhesion	13 proteins
0.000179	26.5	GO:0018298	protein-chromophore linkage	3 proteins
0.000238	8.83	GO:0008544	epidermis development	5 proteins
0.000895	42.9	GO:0007130	synaptonemal complex assembly	2 proteins

Top 5 enriched interactors by enrichment significance (See all here)

p	interactors	enrichment	protein
0.000	5	10.348	Dynein light chain 1, cytoplasmic
0.000	3	28.384	Dynein light chain 2, cytoplasmic
0.001	3	18.062	E3 ubiquitin-protein ligase RFWD2
0.001	2	66.229	Hsp70-binding protein 1
0.001	2	66.229	Dynein light chain Tctex-type 3

Motif Statistics

Type	Motif	Enrichment	Instances	Proteins	Dataset Size
Motif	[KR].TQT	-	161	153	20266
Reverse	TQT.[KR]	0.957	154	152	20266
Shuffle	[KR]TQT.	0.981	158	150	20266

Raw Data

Figure 1. Main results page**.** In addition to individual statistics for the top ten motif occurrences, the default results page displays motif counts, GO enrichment and protein interactor enrichment for all occurrences. If no enriched GO terms and/or interactors have been found, these sections will be blank.

provided. Example proteome restrictions by sequence (the three curated human occurrences of LIG_Dynein_DLC8_1) and GO term (cytoplasmic dynein complex) can also be loaded at the front page.

Getting help

SLiMSearch 2.0 is supported by an extensive help section, including a quickstart guide and walkthrough with screenshots. Example input files are provided and example input data can be loaded into the input forms. Fully interactive example output (corresponding to running the example input with default parameters) is clearly linked from the help pages (See 'Example analysis' section).

FUTURE WORK

Currently, only human proteome searches are available but other proteomes will be added with time. A selection of model organisms will be added in the near future.

CONCLUSION

There are many sources of *de novo* motifs, including experimental approaches such as mutagenesis and peptide arrays or phage display. With recent developments in experimental technologies for determining protein–protein interaction networks and computational techniques for predicting interaction motifs from them, the number of putative SLiMs is likely to increase dramatically in the next few years. SLiMSearch 2.0 represents a valuable tool for the annotation of such motifs. In addition to *de novo* motifs, the server is useful for finding candidate occurrences of established SLiMs, including those found in motif databases such as ELM (2) and MiniMotif Miner (3). Often, the definition of these motifs is not conclusive and so there are also times when it is useful to search using a specific variant or a relaxed motif definition. For many known SLiMs, we currently only have annotated occurrences for a restricted set of taxonomic groups (2) but, due to their short and degenerate nature, they often evolve convergently (31). As the number of full proteomes continues to increase, the SLiMSearch 2.0 server will enable the identification of SLiMs in new taxa, helping to shed light on the breadth and depth of functional SLiMs. The SLiMSearch 2.0 server is available at: http://bioware.ucd.ie/slimsearch2.html.

FUNDING

Science Foundation Ireland (08/IN.1/B1864) and the University of Southampton; European Molecular Biology Laboratory [EMBL Interdisciplinary Postdoc (EIPOD) fellowship to N.E.D.] Funding for open access charge: The University of Southampton.

Conflict of interest statement. None declared.

REFERENCES

1. Diella,F., Haslam,N., Chica,C., Budd,A., Michael,S., Brown,N.P., Trave,G. and Gibson,T.J. (2008) Understanding eukaryotic linear motifs and their role in cell signaling and regulation. *Front. Biosci.*, **13**, 6580–6603.
2. Gould,C.M., Diella,F., Via,A., Puntervoll,P., Gemund,C., Chabanis-Davidson,S., Michael,S., Sayadi,A., Bryne,J.C., Chica,C. *et al.* (2010) ELM: the status of the 2010 eukaryotic linear motif resource. *Nucleic Acids Res.*, **38**, D167–D180.
3. Rajasekaran,S., Balla,S., Gradie,P., Gryk,M.R., Kadaveru,K., Kundeti,V., Maciejewski,M.W., Mi,T., Rubino,N., Vyas,J. *et al.*

(2009) Minimotif miner 2nd release: a database and web system for motif search. *Nucleic Acids Res.*, **37**, D185–D190.
4. Ramu,C. (2003) 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.
5. de Castro,E., Sigrist,C.J., Gattiker,A., Bulliard,V., Langendijk-Genevaux,P.S., Gasteiger,E., Bairoch,A. and Hulo,N. (2006) ScanProsite: detection of PROSITE signature matches and ProRule-associated functional and structural residues in proteins. *Nucleic Acids Res.*, **34**, W362–W365.
6. Gutman,R., Berezin,C., Wollman,R., Rosenberg,Y. and Ben-Tal,N. (2005) QuasiMotiFinder: protein annotation by searching for evolutionarily conserved motif-like patterns. *Nucleic Acids Res.*, **33**, W255–W261.
7. Bateman,A., Coin,L., Durbin,R., Finn,R.D., Hollich,V., Griffiths-Jones,S., Khanna,A., Marshall,M., Moxon,S., Sonnhammer,E.L. *et al.* (2004) The Pfam protein families database. *Nucleic Acids Res.*, **32**, D138–D141.
8. Sigrist,C.J., Cerutti,L., de Castro,E., Langendijk-Genevaux,P.S., Bulliard,V., Bairoch,A. and Hulo,N. (2010) PROSITE, a protein domain database for functional characterization and annotation. *Nucleic Acids Res.*, **38**, D161–D166.
9. Chica,C., Labarga,A., Gould,C.M., Lopez,R. and Gibson,T.J. (2008) A tree-based conservation scoring method for short linear motifs in multiple alignments of protein sequences. *BMC Bioinformatics*, **9**, 229.
10. Davey,N.E., Shields,D.C. and Edwards,R.J. (2009) Masking residues using context-specific evolutionary conservation significantly improves short linear motif discovery. *Bioinformatics*, **25**, 443–450.
11. Via,A., Gould,C.M., Gemund,C., Gibson,T.J. and Helmer-Citterich,M. (2009) A structure filter for the Eukaryotic Linear Motif Resource. *BMC Bioinformatics*, **10**, 351.
12. Jonassen,I., Collins,J.F. and Higgins,D.G. (1995) Finding flexible patterns in unaligned protein sequences. *Protein Sci.*, **4**, 1587–1595.
13. Bailey,T.L., Boden,M., Buske,F.A., Frith,M., Grant,C.E., Clementi,L., Ren,J., Li,W.W. and Noble,W.S. (2009) MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res.*, **37**, W202–W208.
14. Neduva,V. and Russell,R.B. (2006) DILIMOT: discovery of linear motifs in proteins. *Nucleic Acids Res.*, **34**, W350–W355.
15. 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.
16. Edwards,R.J., Davey,N.E. and Shields,D.C. (2007) SLiMFinder: A Probabilistic Method for Identifying Over-Represented, Convergently Evolved, Short Linear Motifs in Proteins. *PLoS ONE*, **2**, e967.
17. Lieber,D.S., Elemento,O. and Tavazoie,S. (2010) Large-scale discovery and characterization of protein regulatory motifs in eukaryotes. *PLoS ONE*, **5**, e14444.
18. Edwards,R.J., Davey,N.E. and Shields,D.C. (2008) CompariMotif: quick and easy comparisons of sequence motifs. *Bioinformatics*, **24**, 1307–1309.
19. Davey,N.E., Haslam,N.J., Shields,D.C. and Edwards,R.J. (2010) SLiMFinder: a web server to find novel, significantly over-represented, short protein motifs. *Nucleic Acids Res.*, **38**, W534–W539.
20. Davey,N.E., Haslam,N.J., Shields,D.C. and Edwards,R.J. (2010) SLiMSearch: a webserver for finding novel occurrences of short linear motifs in proteins, incorporating sequence context. In Dijkstra,T.M.H., Tsivtsivadze,E., Marchiori,E. and Heskes,T. (eds), *Pattern Recognition in Bioinformatics, Lecture Notes in Bioinformatics*, Vol. 6282. Springer, Berlin, pp. 50–61.
21. 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.
22. Fuxreiter,M., Tompa,P. and Simon,I. (2007) Local structural disorder imparts plasticity on linear motifs. *Bioinformatics*, **23**, 950–956.
23. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–29.
24. Stark,C., Breitkreutz,B.J., Chatr-Aryamontri,A., Boucher,L., Oughtred,R., Livstone,M.S., Nixon,J., Van Auken,K., Wang,X., Shi,X. *et al.* (2011) The BioGRID Interaction Database: 2011 update. *Nucleic Acids Res.*, **39**, D698–D704.
25. Davey,N.E., Edwards,R.J. and Shields,D.C. (2007) The SLiMDisc server: short, linear motif discovery in proteins. *Nucleic Acids Res.*, **35**, W455–W459.
26. The UniProt Consortium. (2011) Ongoing and future developments at the Universal Protein Resource. *Nucleic Acids Res.*, **39**, D214–D219.
27. Flicek,P., Amode,M.R., Barrell,D., Beal,K., Brent,S., Chen,Y., Clapham,P., Coates,G., Fairley,S., Fitzgerald,S. *et al.* (2011) Ensembl 2011. *Nucleic Acids Res.*, **39**, D800–D806.
28. Katoh,K. and Toh,H. (2008) Recent developments in the MAFFT multiple sequence alignment program. *Brief. Bioinform.*, **9**, 286–298.
29. Finn,R.D., Mistry,J., Tate,J., Coggill,P., Heger,A., Pollington,J.E., Gavin,O.L., Gunasekaran,P., Ceric,G., Forslund,K. *et al.* (2010) The Pfam protein families database. *Nucleic Acids Res.*, **38**, D211–D222.
30. Rose,P.W., Beran,B., Bi,C., Bluhm,W.F., Dimitropoulos,D., Goodsell,D.S., Prlic,A., Quesada,M., Quinn,G.B., Westbrook,J.D. *et al.* (2011) The RCSB Protein Data Bank: redesigned web site and web services. *Nucleic Acids Res.*, **39**, D392–D401.
31. Davey,N.E., Trave,G. and Gibson,T.J. (2010) How viruses hijack cell regulation. *Trends Biochem. Sci.*, **36**, 159–169.