W464–W469 *Nucleic Acids Research, 2017, Vol. 45, Web Server issue* *Published online 6 April 2017*
doi: 10.1093/nar/gkx238

SLiMSearch: a framework for proteome-wide discovery and annotation of functional modules in intrinsically disordered regions

Izabella Krystkowiak[1,2] and Norman E. Davey[1,2,*]

[1]Conway Institute of Biomolecular & Biomedical Research, University College Dublin, Belfield, Dublin 4, Ireland and [2]UCD School of Medicine & Medical Science, University College Dublin, Belfield, Dublin 4, Ireland

Received January 28, 2017; Revised March 20, 2017; Editorial Decision March 27, 2017; Accepted April 05, 2017

ABSTRACT

The extensive intrinsically disordered regions of higher eukaryotic proteomes contain vast numbers of functional interaction modules known as short linear motifs (SLiMs). Here, we present SLiMSearch, a motif discovery tool that scans a motif consensus, representing the specificity determinants of a motif-binding domain, against a proteome to discover putative novel motif instances. SLiMSearch applies several distinct and complementary approaches exploiting the common properties of SLiMs to predict novel motifs. Consensus matches are annotated with overlapping sequence annotation, including feature information describing protein modular architecture, post-translational modification, structure, sequence variation and experimental characterisation of functional regions. Discriminatory motif attributes such as conservation and accessibility are also calculated. In addition, SLiMSearch provides functional enrichment and evolutionary analysis tools. The enrichment tool analyses GO terms, keywords and interacting partner enrichment to indicate possible motif function. The evolutionary tool evaluates motif taxonomic range and the conservation of motif sequence context. Consensus matches can be filtered based on motif attributes such as accessibility and taxonomic range; or by the localisation, interacting partners or ontology annotation of the peptide-containing protein. SLiMSearch supports a range of species of experimental and therapeutic relevance and is available online at http://slim.ucd.ie/slimsearch/.

INTRODUCTION

The higher eukaryotic proteomes contain extensive intrinsically disordered regions and these regions mediate many of the interactions of a protein (1–4). Furthermore, they integrate information encoded in their environment to make regulatory decisions in reaction to cell state changes (5,6). Various estimates have suggested that there may be upwards of one hundred thousand interaction interfaces in the intrinsically disordered regions of the human proteome (3). Yet, only a small portion of the functional elements predicted to reside within these regions have been characterized (3,7). The majority of experimentally characterized modules in disordered regions belong to a class of compact, degenerate and *ex nihilo* evolvable interaction interfaces known as short, linear motifs (SLiMs) (8–10). SLiMs perform many of the regulatory functions associated with intrinsically disordered regions. They are particularly important for the formation of transient protein complexes, modulating protein modification state, controlling protein stability and directing protein subcellular trafficking (4).

The vast majority of protein motifs remain undiscovered due to experimental and computationally difficulties in characterizing novel motifs (7). Most SLiMs are encoded in a linear region of less than ten amino acids with only three or four core residues determining the majority of the binding affinity and specificity. These core residues extensively contact the motif-binding pocket, and therefore need to be physicochemically compatible with the binding-pocket. This constrains the peptide to a limited set of residues at these positions, resulting in a common motif or consensus in the binding partners of the motif-binding pocket. Consensus searches can be used to discover novel functional SLiMs but their length and the limited number of defined residues makes SLiMs difficult to identify; peptides matching the consensus are very likely to occur by chance, thus, the results are dominated by stochastically occurring non-functional consensus matches (9). The key steps in motif discovery are removing matches that are unlikely to be functional, and annotating the remaining matches with discriminatory data that can be used to prioritize these matches for further experimental validation. In recent work, we leveraged *in silico* sequence analysis to discover and annotate peptides matching the known specificity determinants of two motif-

*To whom correspondence should be addressed. Tel: +353 1 716 6700; Fax: +353 1 716 6701; Email: norman.davey@ucd.ie

© The Author(s) 2017. Published by Oxford University Press on behalf of Nucleic Acids Research.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

binding proteins, the APC/C substrate recruitment subunit Cdc20 and the protein phosphatase $PP2A^{B56}$ holoenzyme (11–13). Discriminatory attributes indicative of motif functionality were used to guide the experimental characterization of several novel motifs, thereby, advocating the use of sequence analysis tools to augment experimental motif discovery.

Several web-based tools are currently available for the discovery of novel instances of SLiM classes with characterized specificity determinants (14,15) (See Supplementary Table S1 for a detailed list). SLiM instance discovery webservers can be split into methods that scan a single protein with a set of predefined functionally characterized motif consensuses such as ELM (16), QuasiMotiFinder (17) and MiniMotifMiner (18); and those that scan a large set of proteins with a single motif consensus such as SLiMSearch (19), ScanProsite (20), SIRW (21), iELM (22) and DoReMi (23). These tools utilize a range of discriminatory attributes to prioritize consensus matches including sequence context, match conservation, structural context, ontology and interaction data to optimize motif discovery through filtering and ranking. Here, we introduce a major update to SLiMSearch (19), a web-based tool for the discovery of novel SLiM instances in a proteome. For this release, SLiMSearch has been completely rewritten from top to bottom. A new data management framework allowing automated dataset construction built on a relational database has replaced the previous flat file data storage framework. Novel conservation, functional analysis and filtering functionality have been added allowing complex querying and filtering options. In addition, the current version has been expanded from a single human dataset to 70 species of experimental and therapeutic relevance. SLiMSearch 4.0 is a single web-based framework that consolidates a decade of research into the discriminatory attributes pertinent to motif discovery. The resulting tool produces an intuitive, informative and interactive output that can be used to identify putative functional modules in the disordered regions of a proteome.

MATERIALS AND METHODS

The SLiMSearch framework is a suite of sequence and data analysis tools for motif consensus search, annotation and filtering. The framework takes as input a motif consensus describing the specificity determinants of a motif-binding pocket in regular expression syntax (see Supplementary Material), and a species of interest. This motif consensus can be derived from experimental data such as peptide mutagenesis, peptide arrays, phage display, motif evolutionary analyses or motif structural characterization (7,24) (see Supplementary Tables S2 and S3). The consensus is scanned against the proteome of the chosen species and a sortable list of annotated consensus matches is returned. These consensus matches can be analysed for evolutionary attributes or functional enrichment. Finally, matches can be filtered based on a wide range of discriminatory attributes. The framework was designed to facilitate easy expansion and updating of the underlying data. This allows novel protein sets, source of discriminatory data and sequence attributes to be effortlessly added to the resource. The current implementation covers 69 species including most major model organisms and relevant pathogens; and has a single protein set covering viral proteins (see Supplementary Table S4). An extensive help page including the required input and a detailed description of the output is available on the website. Jobs are stored on the server for 14 days after which they are deleted.

Feature and attribute annotation

The framework accesses a large pre-computed database of protein-centric information to annotate consensus matches with attributes that are strong discriminators for or against motif functionality (see Supplementary Table S5). Furthermore, consensus matches are annotated with information to understand the pre- and post-translational mechanisms controlling their function (5,6,25). For example, SLiMSearch annotates the motif instances overlapping features describing the protein modular architecture such as short linear motifs and domains (8,26), sites of post-translational modifications (27,28) and protein topology information (29); experimentally characterized regions such as solved structures (30), secondary structure assignment (31), mutagenesis, regions of interest and binding sites (32); and sequence variation such as alternative transcription, alternative splicing and SNPs (33,34). Peptide attributes are also quantified: scored as peptide disorder propensity (35), solvent accessibility (31) (if an overlapping structure is available), and conservation (36) (see Supplementary Material for details). Furthermore, proteins and regions of proteins that are inaccessible to intracellular proteins (e.g. secreted proteins, extracellular protein regions or transmembrane regions) are also annotated. All features and attribute annotation can be used to sort the consensus matches. Each annotation is linked to the source data to obtain more details about a feature or attribute of interest. Finally, the consensus matches are linked, by clicking on the peptide or conservation score, to the ProViz protein visualization tool (http://proviz.ucd.ie) allowing the overlapping feature and attribute annotations to be visualized (37).

Evolutionary annotation

There are two major evolutionary discriminators for motif functionality (Figure 1A): conservation over large evolutionary distances (Figure 1B) and high levels of conservation relative to the flanking regions (Figure 1C and D) (36,38–40). SLiMSearch provides conservation metrics to describe these discriminatory attributes. The taxonomic range section provides information about the conservation of the consensus across a set of species. For each species, a consensus match is annotated regarding its presence or absence at the same position in an ortholog alignment (Figure 1A). Conservation of a motif consensus over a large taxonomic range is a pointer towards a region that is constrained and therefore functional. Hence, experimentally characterized functional motifs are conserved over larger taxonomic ranges than uncharacterized consensus matches (where the majority of instances are non-functional) (Figure 1B). The flank conservation annotates the conservation of the consensus match relative to the conservation

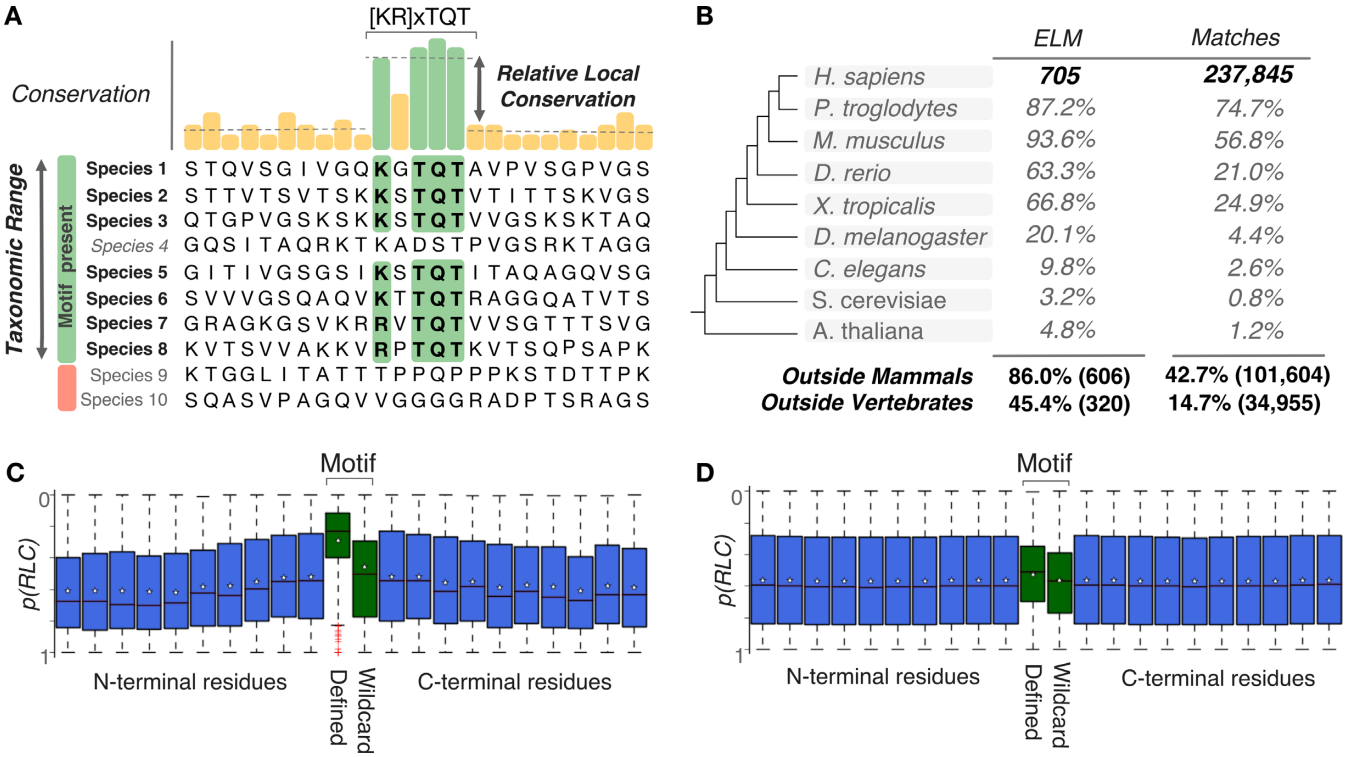

Figure 1. Benchmarking of evolutionary annotation used in SLiMSearch. (**A**) Example alignment of a [KR]xTQT Dynein Light Chain binding motif across different species showing the attributes of motif conservation measured by the relative local conservation and taxonomic range. (**B**) Motif consensus conservation of human motif instances across different species. The motif consensus taxonomic range of the validated human instances in the ELM resource compared to the non-validated instances (Instances in the human proteome which match a motif consensus from the ELM database, but are not annotated as a 'true positive' in the ELM database). (**C**) Relative local conservation (see Supplementary Material) for each residue in the defined, wildcard and flanking regions of a motif for validated instances from the ELM resource. (**D**) Relative local conservation for each residue in the defined, wildcard and flanking regions of a motif for consensus matches not annotated as validated instances from the ELM resource.

of the directly flanking regions and the relative conservation scores quantifies the likelihood of this relative conservation statistically. Similar to taxonomic range, relative conservation is a strong discriminator of functional motifs. The defined residues of functional motifs that make a direct contact with the binding partner are generally more conserved than solvent facing positions and flanking residues (Figure 1C and D). Consequently, in rapidly evolving intrinsically disordered regions, functional motifs are often observed as islands of conservation in a sea of mutations, insertions and deletions. Relative conservation quantifies this property. The flank conservation section also graphically represents the conservation of the sequence context of the consensus match where the level of conservation for each residue within and flanking the match correlates with the colour intensity. All conservation metrics are built on pre-computed ortholog alignments allowing complex evolutionary information to be rapidly computed and accessed (see Supplementary Materials). The alignment of the region used for the conservation calculation can be directly visualized using the ProViz protein visualization tool by clicking on the peptide (37).

Functional enrichment analysis

A set of motif instances recognized by a given motif-binding partner will by definition share a common interactor, however, they often also share a common function, pathway or localisation (8). SLiMSearch analyses the enrichment of GO terms, keywords and interactors for the set of motif-containing proteins to link a function, localisation or binding partner with a motif consensus. Analysis of the ontologies can allow functions to be established for newly discovered motif classes or allow novel aspects of the biology of a previously characterized motif to be uncovered. Functional analysis of motif consensus search data has two biases that render the result of classical hypergeometric-based enrichment analyses unreliable. Firstly, the probability of seeing a consensus match in a protein is correlated to the length of the amino acid sequence. Therefore, functional annotations that are associated with longer proteins are more likely to be significantly enriched. This can be clearly seen when the median enrichment score (See Supplementary Material) of a GO term is plotted against the average number of disordered amino acids per protein annotated with that GO term for random motif consensus test sets (randomized motif consensuses would not be expected to have enriched functional annotation) (Figure 2A). This bias results in strong over-estimates of the significance of certain GO terms and, as a result, even random motif consensuses will regularly have numerous significantly enriched GO terms (Figure 2B). Secondly, related proteins, due to their sequence similarity, are more likely to share a randomly occurring consensus match. However, related proteins are also more likely

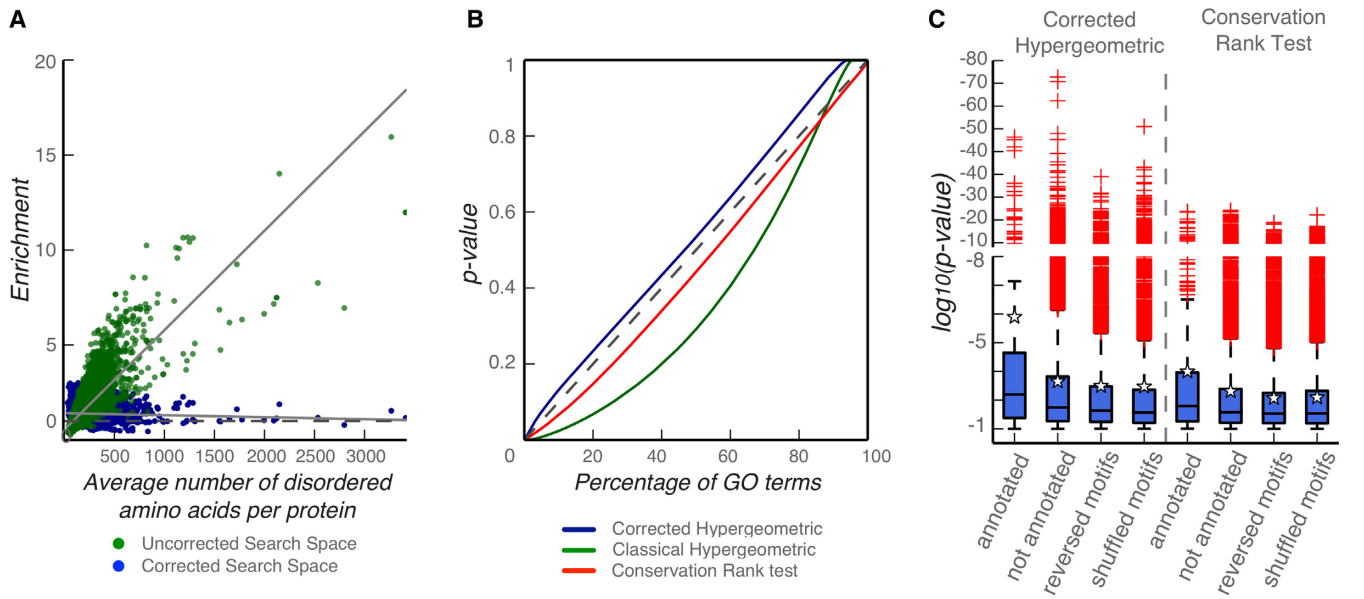

Figure 2. Benchmarking of the functional enrichment analysis approaches used by SLiMSearch. (**A**) Plot of the median GO term enrichment scores against the average number of disordered amino acids per protein for GO terms returned from the enrichment analysis of the random benchmarking set (see Supplementary Material). (**B**) Plot of the average p-value for a GO term against the percentage of GO terms with that p-value or less for the random benchmarking set. In this dataset, which should have no functional motif consensuses and therefore no enriched GO-terms, the data points should fall along the diagonal. The classical hypergeometric test clearly diverges from the diagonal and is under the line, as such it strongly over predicts the significance of each GO term. *P*-values are calculated using classical hypergeometric test with Benjamini–Hochberg correction (classical hypergeometric); hypergeometric test with Benjamini–Hochberg correction with motif search space correction (corrected hypergeometric); and Mann–Whitney U rank test for enrichment analysis based on conservation (QFO) (conservation rank test). (**C**) The distribution of corrected hypergeometric and conservation rank test *P*-values of GO terms for consensus searches of ELM class regular expressions (split into extended GO terms annotated in the ELM resource as functionally related to an ELM class, and extended GO terms not annotated for the ELM class), reversed ELM classes regular expressions and shuffled ELM classes regular expressions. Enrichment analysis performed with motif search space correction (corrected hypergeometric) and based on QFO conservation (conservation rank test). Both analyses used UniRef50 clustering of related proteins. The stars denote the mean value and red plus values denote outliers.

to have overlapping functional annotations. Consequently, functional annotations associated with large protein families are also often significantly enriched. SLiMSearch includes two functional enrichment tools designed to remove these biases from functional analyses of consensus search results (see Supplementary Materials). The first is a corrected hypergeometric test that accounts for motif search space and evolutionary relationships between proteins, and applies Benjamini–Hochberg correction for multiple testing. The second is a Mann–Whitney rank test analysis using relative conservation scores as the ranking criteria. As functional matches of a motif consensus are generally more conserved than stochastically occurring non-functional instances (Figure 1C and D), biologically relevant functional annotations related to the motif consensus will be enriched for highly conserved motif instances, the non-random nature of this distribution can be captured by the rank test (see Supplementary Figure S3). These novel functional enrichment tools are a clear improvement on the commonly used hypergeometric statistic for functional enrichment analysis and conform closely to the expected distribution for random motif consensus test sets (Figure 2B). Furthermore, when analyzing consensus matches of an experimentally characterized motif family the functional enrichment tools can correctly return functional annotation associated with the motif family (Figure 2C).

Filtering

Accessibility is a key discriminatory attribute for motif functionality (9,41,42). Consequently, by default the protein search space is restricted to the intrinsically disordered regions of the proteome (as defined by IUPred with a cut-off of 0.4) though this can be modified on the input page. Further accessibility filtering options include surface accessibility (when a structure is available) and, overlap with Pfam domains, topology and localization. Consensus matches falling within these regions are not filtered automatically as the filters are quite coarse and can remove many functional motif instances. For example, surface accessibility filtering can remove motifs solved while bound to the motif-binding pocket; many Pfam domains contain motifs in accessible loops or are family descriptors for conserved disordered regions; and topology and localisation requirements vary depending on the motif class searched. Consequently, by default, consensus matches that are found within these regions are retained, however, they are flagged in the output and can be removed using the quick filtering options in the top right corner of the instances table. SLiMSearch also allows consensus matches to be filtered based on general motif attributes such as motif taxonomic range; based on specific information about the motif-binding partner such as interactors, function or co-localisation; or simply using a list of GO term or protein accessions. The filtering options allow the user to create a biologically relevant sub-

set of the consensus matches. For example, SLiMSearch can find Type I WW domain consensus matches ([LP]PxY): in intrinsically disordered and intracellular portions of a protein; in the *Caenorhabditis elegans* proteome that are also conserved in *Drosophila melanogaster*; in proteins annotated as 'transcription' or 'hippo signaling'; in a protein that is known to interact with a protein containing a WW domain; or in a protein that shares at least one GO term with the WW domain-containing Yes-associated protein 1 (YAP1). Filtering options can also be chained to allow complex queries to be performed. For example, SLiMSearch can return all human PIP box motif consensus matches (Qxx[IL]xx[FHY][FHY]) found in a nuclear protein, annotated with the keyword 'DNA repair', conserved outside mammals, that occur in previously characterized PCNA interactors.

DISCUSSION

SLiMSearch is an interactive and information-rich yet intuitive motif discovery tool, accessible through a simple motif search interface. The framework searches characterized motif specificity determinants to identify putative novel motif instances. Instances are annotated with accessibility information, evolutionary attributes and experimental information to simplify the process of selecting instances for further validation. As such, SLiMSearch is a powerful tool to aid biologists in building hypotheses and designing experiments by simplifying the analysis of the functional and evolutionary features of motifs (7).

AVAILABILITY

SLiMSearch is available at http://slim.ucd.ie/slimsearch/.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank our collaborators and colleagues for their testing of the SLiMSearch tool. We thank Aino Jarvelin and Richard Edwards for fruitful discussions and critically reading the manuscript.

FUNDING

This work was funded by a Science Foundation Ireland Starting Investigator Research Grant [13/SIRG/2193 to I.K. and N.E.D.]. Funding for open access charge: Science Foundation Ireland.

Conflict of interest statement. None declared.

REFERENCES

1. Tompa,P. (2012) Intrinsically disordered proteins: a 10-year recap. *Trends Biochem. Sci.*, **37**, 509–516.
2. Wright,P.E. and Dyson,H.J. (2015) Intrinsically disordered proteins in cellular signalling and regulation. *Nat. Rev. Mol. Cell Biol.*, **16**, 18–29.
3. Tompa,P., Davey,N.E., Gibson,T.J. and Babu,M.M. (2014) A million peptide motifs for the molecular biologist. *Mol. cell*, **55**, 161–169.
4. Van Roey,K., Uyar,B., Weatheritt,R.J., Dinkel,H., Seiler,M., Budd,A., Gibson,T.J. and Davey,N.E. (2014) Short linear motifs: ubiquitous and functionally diverse protein interaction modules directing cell regulation. *Chem. Rev.*, **114**, 6733–6778.
5. Van Roey,K., Dinkel,H., Weatheritt,R.J., Gibson,T.J. and Davey,N.E. (2013) The switches.ELM resource: a compendium of conditional regulatory interaction interfaces. *Sci. Signal.*, **6**, rs7.
6. Van Roey,K., Gibson,T.J. and Davey,N.E. (2012) Motif switches: decision-making in cell regulation. *Curr. Opin. Struct. Biol.*, **22**, 378–385.
7. Gibson,T.J., Dinkel,H., Van Roey,K. and Diella,F. (2015) Experimental detection of short regulatory motifs in eukaryotic proteins: tips for good practice as well as for bad. *Cell Commun. Signal.*, **13**, 42.
8. Dinkel,H., Van Roey,K., Michael,S., Kumar,M., Uyar,B., Altenberg,B., Milchevskaya,V., Schneider,M., Kuhn,H., Behrendt,A. *et al.* (2016) ELM 2016–data update and new functionality of the eukaryotic linear motif resource. *Nucleic Acids Res.*, **44**, D294–D300.
9. Davey,N.E., Van Roey,K., Weatheritt,R.J., Toedt,G., Uyar,B., Altenberg,B., Budd,A., Diella,F., Dinkel,H. and Gibson,T.J. (2012) Attributes of short linear motifs. *Mol. bioSyst.*, **8**, 268–281.
10. Davey,N.E., Cyert,M.S. and Moses,A.M. (2015) Short linear motifs—ex nihilo evolution of protein regulation. *Cell Commun. Signal.*, **13**, 43.
11. Di Fiore,B., Davey,N.E., Hagting,A., Izawa,D., Mansfeld,J., Gibson,T.J. and Pines,J. (2015) The ABBA motif binds APC/C activators and is shared by APC/C substrates and regulators. *Dev. Cell*, **32**, 358–372.
12. Lu,D., Hsiao,J.Y., Davey,N.E., Van Voorhis,V.A., Foster,S.A., Tang,C. and Morgan,D.O. (2014) Multiple mechanisms determine the order of APC/C substrate degradation in mitosis. *J. Cell Biol.*, **207**, 23–39.
13. Hertz,E.P., Kruse,T., Davey,N.E., Lopez-Mendez,B., Sigurethsson,J.O., Montoya,G., Olsen,J.V. and Nilsson,J. (2016) A conserved motif provides binding specificity to the PP2A-B56 phosphatase. *Mol. Cell*, **63**, 686–695.
14. Bhowmick,P., Guharoy,M. and Tompa,P. (2015) Bioinformatics approaches for predicting disordered protein motifs. *Adv. Exp. Med. Biol.*, **870**, 291–318.
15. Edwards,R.J. and Palopoli,N. (2015) Computational prediction of short linear motifs from protein sequences. *Methods Mol. Biol.*, **1268**, 89–141.
16. Dinkel,H., Michael,S., Weatheritt,R.J., Davey,N.E., Van Roey,K., Altenberg,B., Toedt,G., Uyar,B., Seiler,M., Budd,A. *et al.* (2012) ELM–the database of eukaryotic linear motifs. *Nucleic Acids Res.*, **40**, D242–D251.
17. 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.
18. Mi,T., Merlin,J.C., Deverasetty,S., Gryk,M.R., Bill,T.J., Brooks,A.W., Lee,L.Y., Rathnayake,V., Ross,C.A., Sargeant,D.P. *et al.* (2012) Minimotif Miner 3.0: database expansion and significantly improved reduction of false-positive predictions from consensus sequences. *Nucleic Acids Res.*, **40**, D252–D260.
19. Davey,N.E., Haslam,N.J., Shields,D.C. and Edwards,R.J. (2011) SLiMSearch 2.0: biological context for short linear motifs in proteins. *Nucleic Acids Res.*, **39**, W56–W60.
20. 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.
21. 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.
22. Weatheritt,R.J., Luck,K., Petsalaki,E., Davey,N.E. and Gibson,T.J. (2012) The identification of short linear motif-mediated interfaces within the human interactome. *Bioinformatics (Oxford, England)*, **28**, 976–982.
23. Horn,H., Haslam,N. and Jensen,L.J. (2014) DoReMi: context-based prioritization of linear motif matches. *PeerJ*, **2**, e315.

24. Blikstad,C. and Ivarsson,Y. (2015) High-throughput methods for identification of protein-protein interactions involving short linear motifs. *Cell Commun. Signal.*, **13**, 38.
25. Weatheritt,R.J., Davey,N.E. and Gibson,T.J. (2012) Linear motifs confer functional diversity onto splice variants. *Nucleic Acids Res.*, **40**, 7123–7131.
26. Finn,R.D., Coggill,P., Eberhardt,R.Y., Eddy,S.R., Mistry,J., Mitchell,A.L., Potter,S.C., Punta,M., Qureshi,M., Sangrador-Vegas,A. *et al.* (2016) The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res.*, **44**, D279–D285.
27. Hornbeck,P.V., Kornhauser,J.M., Tkachev,S., Zhang,B., Skrzypek,E., Murray,B., Latham,V. and Sullivan,M. (2012) PhosphoSitePlus: a comprehensive resource for investigating the structure and function of experimentally determined post-translational modifications in man and mouse. *Nucleic Acids Res.*, **40**, D261–D270.
28. Dinkel,H., Chica,C., Via,A., Gould,C.M., Jensen,L.J., Gibson,T.J. and Diella,F. (2011) Phospho.ELM: a database of phosphorylation sites–update 2011. *Nucleic Acids Res.*, **39**, D261–D267.
29. Krogh,A., Larsson,B., von Heijne,G. and Sonnhammer,E.L. (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol.*, **305**, 567–580.
30. Touw,W.G., Baakman,C., Black,J., te Beek,T.A., Krieger,E., Joosten,R.P. and Vriend,G. (2015) A series of PDB-related databanks for everyday needs. *Nucleic Acids Res.*, **43**, D364–D368.
31. Kabsch,W. and Sander,C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577–2637.
32. UniProt,C. (2015) UniProt: a hub for protein information. *Nucleic Acids Res.*, **43**, D204–D212.
33. Sherry,S.T., Ward,M.H., Kholodov,M., Baker,J., Phan,L., Smigielski,E.M. and Sirotkin,K. (2001) dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.*, **29**, 308–311.
34. Abecasis,G.R., Auton,A., Brooks,L.D., DePristo,M.A., Durbin,R.M., Handsaker,R.E., Kang,H.M., Marth,G.T., McVean,G.A. and 1000 Genomes Project, C. (2012) An integrated map of genetic variation from 1,092 human genomes. *Nature*, **491**, 56–65.
35. Dosztanyi,Z., Csizmok,V., Tompa,P. and Simon,I. (2005) The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins. *J. Mol. Biol.*, **347**, 827–839.
36. Davey,N.E., Cowan,J.L., Shields,D.C., Gibson,T.J., Coldwell,M.J. and Edwards,R.J. (2012) SLiMPrints: conservation-based discovery of functional motif fingerprints in intrinsically disordered protein regions. *Nucleic Acids Res.*, **40**, 10628–10641.
37. Jehl,P., Manguy,J., Shields,D.C., Higgins,D.G. and Davey,N.E. (2016) ProViz-a web-based visualization tool to investigate the functional and evolutionary features of protein sequences. *Nucleic Acids Res.*, **44**, W11–W15.
38. 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.
39. 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 (Oxford, England)*, **25**, 443–450.
40. Nguyen Ba,A.N., Yeh,B.J., van Dyk,D., Davidson,A.R., Andrews,B.J., Weiss,E.L. and Moses,A.M. (2012) Proteome-wide discovery of evolutionary conserved sequences in disordered regions. *Science Signal.*, **5**, rs1.
41. Fuxreiter,M., Tompa,P. and Simon,I. (2007) Local structural disorder imparts plasticity on linear motifs. *Bioinformatics*, **23**, 950–956.
42. 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.