doi: 10.1093/nar/gkw265

ProViz—a web-based visualization tool to investigate the functional and evolutionary features of protein sequences

Peter Jehl[1,2,†], Jean Manguy[1,2,†], Denis C. Shields[1,2], Desmond G. Higgins[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 31, 2016; Revised March 27, 2016; Accepted April 4, 2016

ABSTRACT

Low-throughput experiments and high-throughput proteomic and genomic analyses have created enormous quantities of data that can be used to explore protein function and evolution. The ability to consolidate these data into an informative and intuitive format is vital to our capacity to comprehend these distinct but complementary sources of information. However, existing tools to visualize protein-related data are restricted by their presentation, sources of information, functionality or accessibility. We introduce ProViz, a powerful browser-based tool to aid biologists in building hypotheses and designing experiments by simplifying the analysis of functional and evolutionary features of proteins. Feature information is retrieved in an automated manner from resources describing protein modular architecture, post-translational modification, structure, sequence variation and experimental characterization of functional regions. These features are mapped to evolutionary information from precomputed multiple sequence alignments. Data are displayed in an interactive and information-rich yet intuitive visualization, accessible through a simple protein search interface. This allows users with limited bioinformatic skills to rapidly access data pertinent to their research. Visualizations can be further customized with user-defined data either manually or using a REST API. ProViz is available at http://proviz.ucd.ie/.

INTRODUCTION

Proteins are modular entities consisting of autonomous functional regions such as globular domains (1–3), disordered domains (4–6) and short linear motifs (SLiMs) (7–9). These modules are regularly modulated by post-translational modifications (PTMs) (10,11) and can be added or removed by alternative transcription or alternative splicing to produce protein isoforms with unique functional properties (12,13). Furthermore, single nucleotide polymorphisms (SNPs) can disrupt the normal function of these modules often resulting in deleterious outcomes that underlie disease (14,15). Over the past decade advances in biochemical, proteomic and genomic methods have rapidly expanded our understanding of these aspects of protein biology. Biochemical studies have revealed residues and regions of functional importance. Structural biology techniques have produced detailed structures of protein regions both in their unbound and bound state (16). Proteomic studies continue to expand the census of PTMs (10,11) and are now being applied to the difficult task of SLiM discovery (17,18). Genomic studies have catalogued both disease-causing and natural variant non-synonymous SNPs (14,15); and temporal, spatial or cell type-specific regions of proteins encoded by non-constitutive exons (13). Computational methods can accurately predict protein sequence features and attributes. Homology-based inference is widely used to map functional modules to regions of unstudied proteins (19). Sequence analysis tools can accurately define regions of proteins that are unlikely to have any structure in their native state (20,21) and predict protein membrane topology relative to membrane crossing regions (22). Finally, large-scale sequencing efforts have produced complete bacterial, archeal, eukaryotic and viral proteomes allowing detailed investigation of protein sequence evolution, thereby pinpointing regions of functional constraint (23,13).

The accumulation of these data has expanded our understanding of many aspects of protein function. However, we are faced with enormous quantities of data dispersed across many different resources. Tools for the aggregation and visualization of a protein's modular architecture, experimental information and evolution are therefore central to our

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

[†] These authors contributed equally to the paper as first authors.

ability to consolidate and digest this information. Existing tools to visualize protein-related data are restricted by their accessibility, sparse functionality, limited sources of data and lack of user friendly interfaces. Several powerful tools allowing complex alignment and feature manipulation that run locally on a user's machines are available (for example, JalView (24), CLC-Workbench and STRAP (25)). However, they require a knowledge of resources and methods to access pertinent data and their extensive functionality far exceeds the requirements of many users. Recent developments driven by novel JavaScript libraries, HTML5 and CSS3 have expanded the potential capabilities of browser-based bioinformatics tools. For many computationally and data intensive bioinformatics problems browser-based applications are still unsuitable. For certain tasks, however, these developments provide a powerful framework. In particular, the graphical ability and interactivity of such a framework are ideally suited to visualization tools. Although, developers must still be careful as interfaces can become sluggish if large amounts of data are not handled carefully. To date, several browser-based protein data visualization tools of varying degrees of sophistication have been developed including Java applet based tools (JalView2 Lite (24), STRAP (25), PFAAT (26)), resource specific tools (PDB viewers (27,28), Pfam (3), webPrank (29)) and general alignment/feature viewers (Alignment-Annotator (30), MView (31), JSAV (32)) (Supplementary Table S1). However, none of these tools truly leverage the available protein data resources and web frameworks to their fullest potential. To rectify this we have developed ProViz, a novel interactive browser-based visualization tool to investigate the functional and evolutionary features of protein sequences.

MATERIALS AND METHODS

Input options

ProViz provides a simple search interface built on the UniProt protein search engine (12). The search interface takes a gene or protein name and creates a table of results from which the protein of interest can be chosen. In cases where the protein of interest is not returned a search can be refined by adding the species of interest or, if available, the UniProt identifier (e.g SRC_HUMAN) or UniProt accession (e.g. P12931).

Main visualization

The ProViz main visualization displays a protein of interest as a single linear peptide. This query protein is annotated with evolutionary and functional data (Figure 1). The data are mapped to each residue or a range of residues directly below the query sequence. The visualization consists of two sections: sequence data and feature data. The sequence data section displays the sequence of the query protein and, when available, a multiple sequence alignment of proteins homologous to the query protein. The feature data section displays data on the modular architecture, PTM state, structure, sequence variation and experimental characterization of functional regions for the query protein. An additional visualization, the protein architecture section, displays an overview of the query protein annotated with key features.

Sequence data. The sequence data section displays protein and alignment data (Figure 2A). ProViz visualizations are built around a query sequence and, consequently, the query sequence is always displayed by default. Where possible, a multiple sequence alignment of proteins homologous to the query protein is displayed. To enable the mapping of features to the protein sequences, the alignment is degapped with respect to the query sequence. That is, columns of the alignment which are gaps in the query sequence are hidden. Residues flanking these hidden regions are displayed in lowercase. The sequence of a hidden region can be displayed by hovering over the flanking amino acids. A complete gapped alignment can be displayed using the options toolbar, but, feature data are not shown in this view. Both the query protein sequence and alignments are coloured according to the rules of the ClustalX (33) colouring scheme to highlight conserved and physicochemically similar amino acids. ProViz accesses precomputed alignments automatically from two sources: orthologue alignments created by GOPHER (Generation of Orthologous Proteins from High-throughput Estimation of relationships) (34) (Supplementary Table S2) and GeneTree homologue alignments from EnsEMBL (13). GeneTree alignments can be filtered to display paralogue or orthologue alignments as well as complete homologue alignments. The available alignments for a query protein are displayed and can be selected in the options toolbar. Due to browser performance issues large alignments are restricted to model organisms by default, however, aligned proteins or the whole alignment can be added back to the visualization by the user through the options sidebar.

Feature data. The query protein sequence is also annotated with protein feature data from numerous resources (Figure 2A and Supplementary Table S3) (3,9,10,12–15,18,19,22,35,36). The presented feature data describes diverse aspects of protein biology (Figure 2B). Complementary computed results from bioinformatics tools are also presented, including disorder predictions (20,21), binding site predictions (37), SLiM consensus matches (9) and residue relative conservation scores (38,39). Three different classes of feature data tracks are used (Figure 2C). Features mapping to a continuous segment of the query proteins (for example, a domain or transmembrane region) are displayed as horizontal bars spanning the corresponding residues of the proteins. Bars are also used to display single amino acid features e.g. modification sites or SNPs. Peptide tracks are similar to bar tracks but display amino acids directly below the corresponding residue in the query protein. Peptide tracks are used for displaying the exact sequence of a region or amino acid of interest such as the tested residue or residues in a mutagenesis experiment. Histogram tracks display quantitative data for the protein on a residue by residue basis. Data are displayed as vertical blocks corresponding to the value given to the residue. Positive and negative values are possible and values are normalized to fit the track height.

Protein architecture. The protein architecture section displays an overview of the query protein (Figure 2A). The section shows a compact visualization of the protein architec-

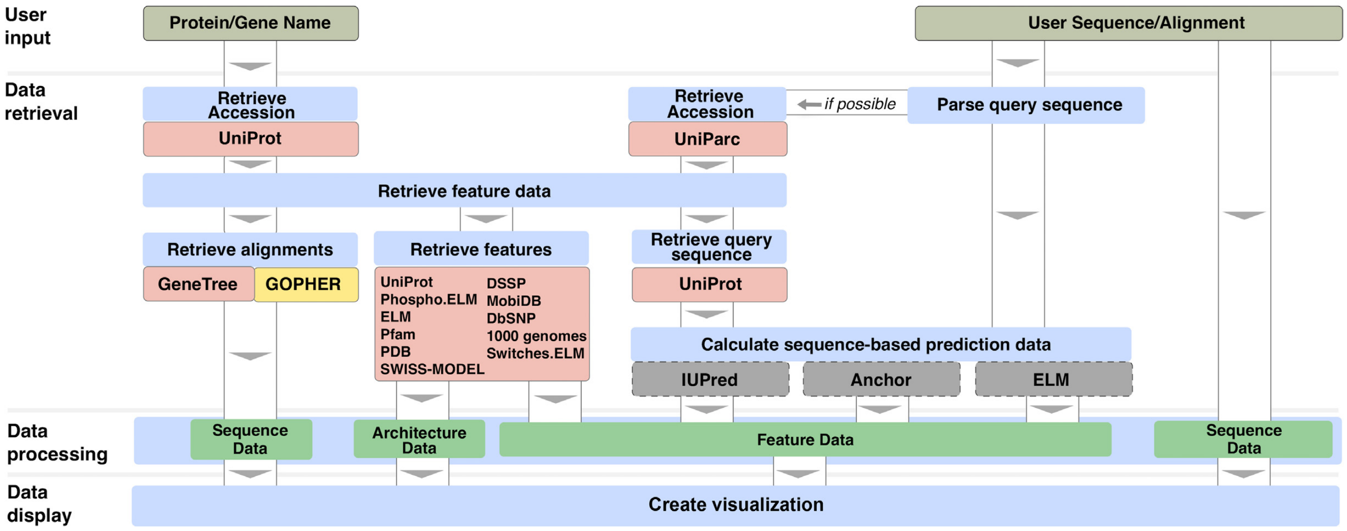

Figure 1. Schema describing data retrieval, data processing and data display for the ProViz protein visualization tool. A user inputs a protein search term or user-defined sequence data (a protein sequence or protein multiple sequence alignment). Search terms (and, if possible, user-defined sequence data) are mapped to a UniProt accession. On the server-side, the UniProt accession is used to retrieve feature data from various resources and sequence-based prediction tools are applied to the protein sequences. All data is processed and returned to the browser-based ProViz front end for visualization. Blue boxes denote functions, red boxes denote external data sources, yellow boxes denote local data sources, grey boxes denote local bioinformatic tools and green boxes denote processed data.

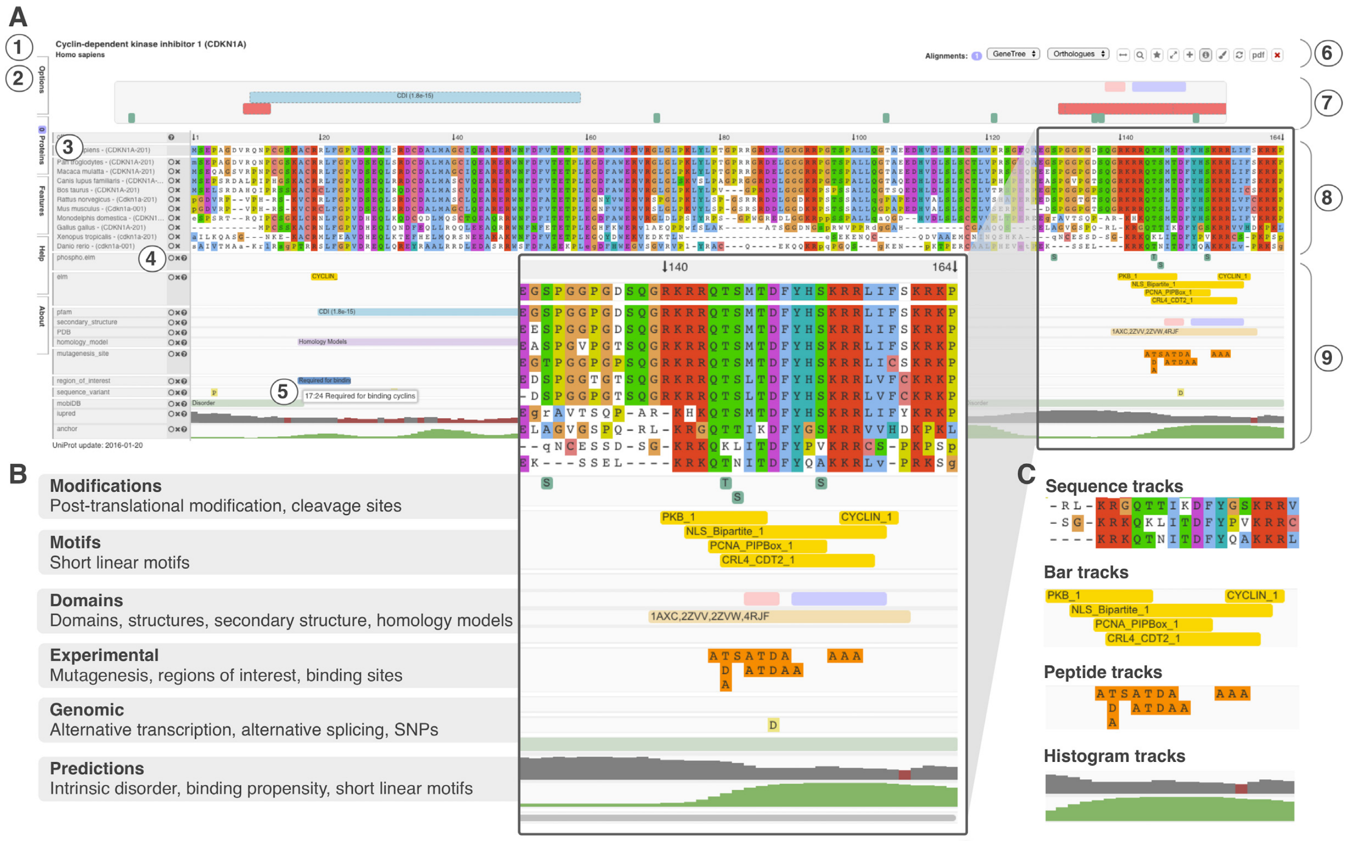
A
1
2
3
4
5
6
7
8
9
Cyclin-dependent kinase inhibitor 1 (CDKN1A)
Homo sapiens
Alignments:
GeneTree
Orthologues
pdf
Options
Proteins
Features
Help
About
CDI (1.6e-15)
17:24 Required for binding cyclins
UniProt update: 2016-01-20
140
164
B
Modifications
Post-translational modification, cleavage sites
Motifs
Short linear motifs
Domains
Domains, structures, secondary structure, homology models
Experimental
Mutagenesis, regions of interest, binding sites
Genomic
Alternative transcription, alternative splicing, SNPs
Predictions
Intrinsic disorder, binding propensity, short linear motifs
PKB_1
CYCLIN_1
NLS_Bipartite_1
PCNA_PIPBox_1
CRL4_CDT2_1
1AXC,2ZVV,2ZVW,4RJF
C
Sequence tracks
Bar tracks
Peptide tracks
Histogram tracks

Figure 2. (**A**) ProViz visualization for Cyclin-dependent kinase inhibitor 1A (CDKN1A) showing selected features of CDKN1A and a GeneTree alignment of CDKN1A orthologues. Key aspects of the visualization are numbered: (1) Protein name and species; (2) options sidebar; (3) data information sidebar; (4) data select, hide and help buttons; (5) information hover tooltip; (6) options toolbar; (7) protein architecture overview; (8) protein sequence data; (9) protein feature data. The visualization in the example can be viewed at http://proviz.ucd.ie/proviz.php?uniprot_acc=P38936. (**B**) A zoomed view of a section of the visualization from panel A labelled with the types of data that are present in each section of the protein feature data. (**C**) Examples of the available track types.

ture showing key features of the protein: secondary structure, topology, globular domains, SLiMs and PTM sites. Users can move a slider or ctrl + click on a feature to rapidly navigate to the specific region of the main visualization. The overview also identifies the region of the query protein currently displayed in the main visualization.

Interactivity. The displayed ProViz data views are highly interactive and customizable. All sequences and tracks can be reordered or hidden. Alignment data can be restricted to proteins from a set of model organisms. Hidden data can be added back to the visualization using the options sidebar. When sequences are removed from an alignment, residues of the reduced alignment can be recoloured and columns solely consisting of gaps can be removed. Users can directly specify a range of residues, move a slider or ctrl + click on a feature to select a target area. The selected area can be used to resize or highlight the visualization. A condensed view is available that shrinks the visualization to double the viewable area. Most elements in the visualization have associated tooltips which upon hovering show detailed information about the element. All features open the web page of their source data upon clicking. Similarly, protein labels link to protein source data. A small tab next to the protein labels opens a new ProViz visualization with the selected protein as the new query sequence. ProViz alignments can be downloaded in FASTA format and the complete visualization can be downloaded in PDF format. Finally, a regular expression can be searched against the sequence data to highlight matching protein subsequences.

Advanced options and customization. ProViz's advanced visualization customization allows user to integrate ProViz into biological resources or present external data. An extensive list of URL options can produce customized visualizations. For example, a specific alignment, set of features or range of amino acids can be be displayed (see Supplementary Table S4 for more details). Custom proteins or alignments can be submitted in FASTA format via the homepage. As ProViz relies on the UniProt accession to retrieve features for the query protein, an MD5 hash value is calculated for the query protein and is searched against the UniParc database. For alignments, the first protein of the alignment is taken as the query protein. If the protein's UniProt identifier is identified, feature data are loaded as normal. If the protein's UniProt identifier cannot be identified, no features are shown. Sequence dependent predictive tools, however, and user provided custom information are still displayed. ProViz also offers the option to add custom tracks to the main visualization. Bars, histogram and peptide tracks are all available. This can be achieved by providing a file in 'XML', 'CSV' or 'JSON' format by drag/drop, file upload or URL based direction to a REST service created or pre-computed file in ProViz format. Example files for custom feature input in 'XML', 'CSV' and 'JSON' formats are available on the ProViz website and are described in the Supplementary Data (Supplementary Tables S5 and S6 and Supplementary File 7).

DISCUSSION

We have introduced ProViz, a novel browser-based interactive exploration tool to investigate the functional and evolutionary features of protein sequences. The browser-based interface of ProViz provides numerous advantages, especially relating to ease of accessibility. However, there are also drawbacks. The most obvious issue is browser performance which can result in a lag when displaying or interacting with visualizations containing large amounts of data. Big alignments and large proteins are a particularly problem and may cause significant problems on older browsers and low specification machines. Similarly, much of the complex functionality that can be performed instantly in local protein data visualization tools, such as recolouring alignments upon sequence removal, require server-side recalculation of the visualization data. As such ProViz should be considered a protein exploration tool that includes protein multiple sequence alignments rather than a fully functioned protein multiple sequence alignment viewer. Nevertheless, ProViz provides a unique resource to quickly gain insight into the function and evolution of proteins. ProViz is a versatile tool that can aid biologist in many ways, for example, building hypotheses, designing experiments or understanding experimental results. The flexible customization options of ProViz also permit bioinformaticians to extensively customize the visualization, for example, by mapping data from high-throughput proteomic studies or adding the results of per residue predictive bioinformatic tools to the presented data. The current version of the tool provides a solid foundation on which to build in the future. Planned extensions include incorporating additional relevant sources of data, sequence-based prediction tools and alignment colouring options. We believe ProViz is an invaluable time-saving resource that will become an integral part of the day to day work of a biologist.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank our collaborators and colleagues for their testing of the ProViz tool. We thank Aino Järvelin, Lisa Rogers and Fabian Sievers for fruitful discussions and critically reading the manuscript.

FUNDING

Science Foundation Ireland Starting Investigator Research Grant [13/SIRG/2193 to N.E.D.]; Irish Research Council (award to Bioinformatics and Systems Biology GREP PhD Programme) (to J.M.); Science Foundation Ireland Principal Investigator grant [08/IN1/B1864 to D.C.S.]; Science Foundation Ireland Principal Investigator grant [11/PI/1034 to D.G.H., P.J.]. Funding for open access charge: Science Foundation Ireland.

Conflict of interest statement. None declared.

REFERENCES

1. Vogel,C., Bashton,M., Kerrison,N.D., Chothia,C. and Teichmann,S.A. (2004) Structure, function and evolution of multidomain proteins. *Curr. Opin. Struct. Biol.*, **14**, 208–216.
2. Han,J.-H., Batey,S., Nickson,A.A., Teichmann,S.A. and Clarke,J. (2007) The folding and evolution of multidomain proteins. *Nat. Rev. Mol. Cell Biol.*, **8**, 319–330.
3. Finn,R.D., Bateman,A., Clements,J., Coggill,P., Eberhardt,R.Y., Eddy,S.R., Heger,A., Hetherington,K., Holm,L., Mistry,J. *et al.* (2014) Pfam: the protein families database. *Nucleic Acids Res.*, **42**, D222–D230.
4. Dyson,H.J. and Wright,P.E. (2005) Intrinsically unstructured proteins and their functions. *Nat. Rev. Mol. Cell Biol.*, **6**, 197–208.
5. Wright,P.E. and Dyson,H.J. (2015) Intrinsically disordered proteins in cellular signalling and regulation. *Nat. Rev. Mol. Cell Biol.*, **16**, 18–29.
6. Guharoy,M., Pauwels,K. and Tompa,P. (2015) SnapShot: intrinsic structural disorder. *Cell*, **161**, 1230.
7. 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.
8. 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.
9. 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.
10. 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.
11. 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.
12. UniProt Consortium. (2015) UniProt: a hub for protein information *Nucleic Acids Res.*, **43**, D204–D212.
13. Yates,A., Akanni,W., Amode,M.R., Barrell,D., Billis,K., Carvalho-Silva,D., Cummins,C., Clapham,P., Fitzgerald,S., Gil,L. *et al.* (2016) Ensembl 2016. *Nucleic Acids Res.*, **44**, D710–D716.
14. 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 Consortium. (2012) An integrated map of genetic variation from 1,092 human genomes. *Nature*, **491**, 56–65.
15. 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.
16. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.
17. 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.
18. Blikstad,C. and Ivarsson,Y. (2015) High-throughput methods for identification of protein-protein interactions involving short linear motifs. *Cell Commun. Signal.*, **13**, 38.
19. Biasini,M., Bienert,S., Waterhouse,A., Arnold,K., Studer,G., Schmidt,T., Kiefer,F., Cassarino,T.G., Bertoni,M., Bordoli,L. *et al.* (2014) SWISS-MODEL: modelling protein tertiary and quaternary structure using evolutionary information. *Nucleic Acids Res.*, **42**, W252–W258.
20. Potenza,E., Di Domenico,T., Walsh,I. and Tosatto,S.C.E. (2015) MobiDB 2.0: an improved database of intrinsically disordered and mobile proteins. *Nucleic Acids Res.*, **43**, D315–D320.
21. 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.
22. 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.
23. Kuzniar,A., van Ham,R.C.H.J., Pongor,S. and Leunissen,J.A.M. (2008) The quest for orthologs: finding the corresponding gene across genomes. *Trends Genet.*, **24**, 539–551.
24. Waterhouse,A.M., Procter,J.B., Martin,D.M.A., Clamp,M. and Barton,G.J. (2009) Jalview Version 2–a multiple sequence alignment editor and analysis workbench. *Bioinformatics*, **25**, 1189–1191.
25. Gille,C. (2006) Structural interpretation of mutations and SNPs using STRAP-NT. *Protein Sci.*, **15**, 208–210.
26. Caffrey,D.R., Dana,P.H., Mathur,V., Ocano,M., Hong,E.-J., Wang,Y.E., Somaroo,S., Caffrey,B.E., Potluri,S. and Huang,E.S. (2007) PFAAT version 2.0: a tool for editing, annotating, and analyzing multiple sequence alignments. *BMC Bioinformatics*, **8**, 381.
27. Moreland,J.L., Gramada,A., Buzko,O. V, Zhang,Q. and Bourne,P.E. (2005) The Molecular Biology Toolkit (MBT): a modular platform for developing molecular visualization applications. *BMC Bioinformatics*, **6**, 21.
28. Rose,P.W., Prlic,A., Bi,C., Bluhm,W.F., Christie,C.H., Dutta,S., Green,R.K., Goodsell,D.S., Westbrook,J.D., Woo,J. *et al.* (2015) The RCSB Protein Data Bank: views of structural biology for basic and applied research and education. *Nucleic Acids Res.*, **43**, D345–D356.
29. Löytynoja,A. and Goldman,N. (2010) webPRANK: a phylogeny-aware multiple sequence aligner with interactive alignment browser. *BMC Bioinformatics*, **11**, 579.
30. Gille,C., Fahling,M., Weyand,B., Wieland,T. and Gille,A. (2014) Alignment-Annotator web server: rendering and annotating sequence alignments. *Nucleic Acids Res.*, **42**, W3–W6.
31. Brown,N.P., Leroy,C. and Sander,C. (1998) MView: a web-compatible database search or multiple alignment viewer. *Bioinformatics*, **14**, 380–381.
32. Martin,A.C.R. (2014) Viewing multiple sequence alignments with the JavaScript Sequence Alignment Viewer (JSAV). *F1000Res.*, **3**, 249.
33. Larkin,M.A., Blackshields,G., Brown,N.P., Chenna,R., McGettigan,P.A., McWilliam,H., Valentin,F., Wallace,I.M., Wilm,A., Lopez,R. *et al.* (2007) Clustal W and Clustal X version 2.0. *Bioinformatics*, **23**, 2947–2948.
34. 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.
35. Touw,W.G., Baakman,C., Black,J., te Beek,T.A.H., Krieger,E., Joosten,R.P. and Vriend,G. (2015) A series of PDB-related databanks for everyday needs. *Nucleic Acids Res.*, **43**, D364–D368.
36. 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.
37. Meszaros,B., Simon,I. and Dosztanyi,Z. (2009) Prediction of protein binding regions in disordered proteins. *PLoS Comput. Biol.*, **5**, e1000376.
38. 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.
39. Sievers,F., Wilm,A., Dineen,D., Gibson,T.J., Karplus,K., Li,W., Lopez,R., McWilliam,H., Remmert,M., Söding,J. *et al.* (2011) Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.*, **7**, 539.