RepeatsDB 2.0: improved annotation, classification, search and visualization of repeat protein structures

Lisanna Paladin, Layla Hirsh, Damiano Piovesan, Miguel A. Andrade-Navarro, Andrey V. Kajava, Silvio C.E. Tosatto

D308–D312 *Nucleic Acids Research, 2017, Vol. 45, Database issue* *Published online 28 November 2016*
doi: 10.1093/nar/gkw1136

RepeatsDB 2.0: improved annotation, classification, search and visualization of repeat protein structures

Lisanna Paladin[1,†], Layla Hirsh[1,2,†], Damiano Piovesan[1], Miguel A. Andrade-Navarro[3], Andrey V. Kajava[4,5,6] and Silvio C.E. Tosatto[1,7,*]

[1]Dept. of Biomedical Sciences, University of Padua, 35121 Padova, Italy, [2]Departamento de Ingeniería, Pontificia Universidad Católica del Perú, 32 Lima, Perú, [3]Institute of Molecular Biology, Faculty of Biology, Johannes Gutenberg University of Mainz, 55128 Mainz, Germany, [4]Centre de Recherches de Biochimie Macromoléculaire, CNRS, Université Montpellier, 34293 Montpellier, France, [5]Institut de Biologie Computationnelle (IBC), 34293 Montpellier, France, [6]Institute of Bioengineering, University ITMO, 197101 St. Petersburg, Russia and [7]CNR Institute of Neuroscience, 35121 Padova, Italy

Received September 15, 2016; Revised October 20, 2016; Editorial Decision October 21, 2016; Accepted October 31, 2016

ABSTRACT

RepeatsDB 2.0 (URL: http://repeatsdb.bio.unipd.it/) is an update of the database of annotated tandem repeat protein structures. Repeat proteins are a widespread class of non-globular proteins carrying heterogeneous functions involved in several diseases. Here we provide a new version of RepeatsDB with an improved classification schema including high quality annotations for ∼5400 protein structures. RepeatsDB 2.0 features information on start and end positions for the repeat regions and units for all entries. The extensive growth of repeat unit characterization was possible by applying the novel ReUPred annotation method over the entire Protein Data Bank, with data quality is guaranteed by an extensive manual validation for >60% of the entries. The updated web interface includes a new search engine for complex queries and a fully re-designed entry page for a better overview of structural data. It is now possible to compare unit positions, together with secondary structure, fold information and Pfam domains. Moreover, a new classification level has been introduced on top of the existing scheme as an independent layer for sequence similarity relationships at 40%, 60% and 90% identity.

INTRODUCTION

Tandem repeat regions in proteins are characterized by a repeated sequence coding for a modular architecture, where structural modules are called 'units'. Proteins with tandem repeats play important functional roles (1), are abundant in nature and related to major health threats (2–6). Detecting and annotating them appropriately may increase our understanding of mechanisms of pathogenicity (e.g. virulence factors (7)), allow the design of scaffold proteins for engineered ligand binding with multiple applications (e.g. cancer therapy (8)) and generally expand our knowledge of the function and structure of many proteins (e.g. the mineralocorticoid receptor (9)). It is widely accepted that structural and functional complexity in domains evolved through fusion, recombination, accretion and repetition of a very small set of elementary functions (10,11). Therefore, units in tandem repeat proteins represent a fundamental source of information to explain contemporary structural diversity and the physico-chemical properties of highly designable folds (12). However, identification of sequence periodicity is an extremely hard task, since repetitive proteins evolve quickly, for two main reasons. The process of duplication originating new repeats is error-prone and identical flanking repeat units have an intrinsic tendency to diverge (13). A number of structure-based methods for the identification of repeats has been developed to fill this gap (14–17). RepeatsDB (18) was proposed in 2014 as a database of repeat protein structures and a resource for high-quality repeat structure annotation. The data was collected with RAPHAEL (19), a state-of-the-art method for the detection of Protein Data Bank (PDB) (20) structures containing repeat regions. The entries were classified into repeat structural classes (21) and further divided into subclasses. The five repeat classes are mainly distinguished by repeat unit length and general structural arrangement, and subclasses by the secondary structure assignment of the repeat unit. The shortest repeats, of one or two residues, form crystallites and are typically harmful or non-functional in natural organisms. No example of their structure is present in the PDB and consequently in RepeatsDB. Class II structures are fibrous proteins with very short units stabilized by interchain interactions, typically

*To whom correspondence should be addressed. Tel: +39 0498276269; Fax: +39 0498276260; Email: silvio.tosatto@unipd.it
†These authors contributed equally to the paper as first authors.

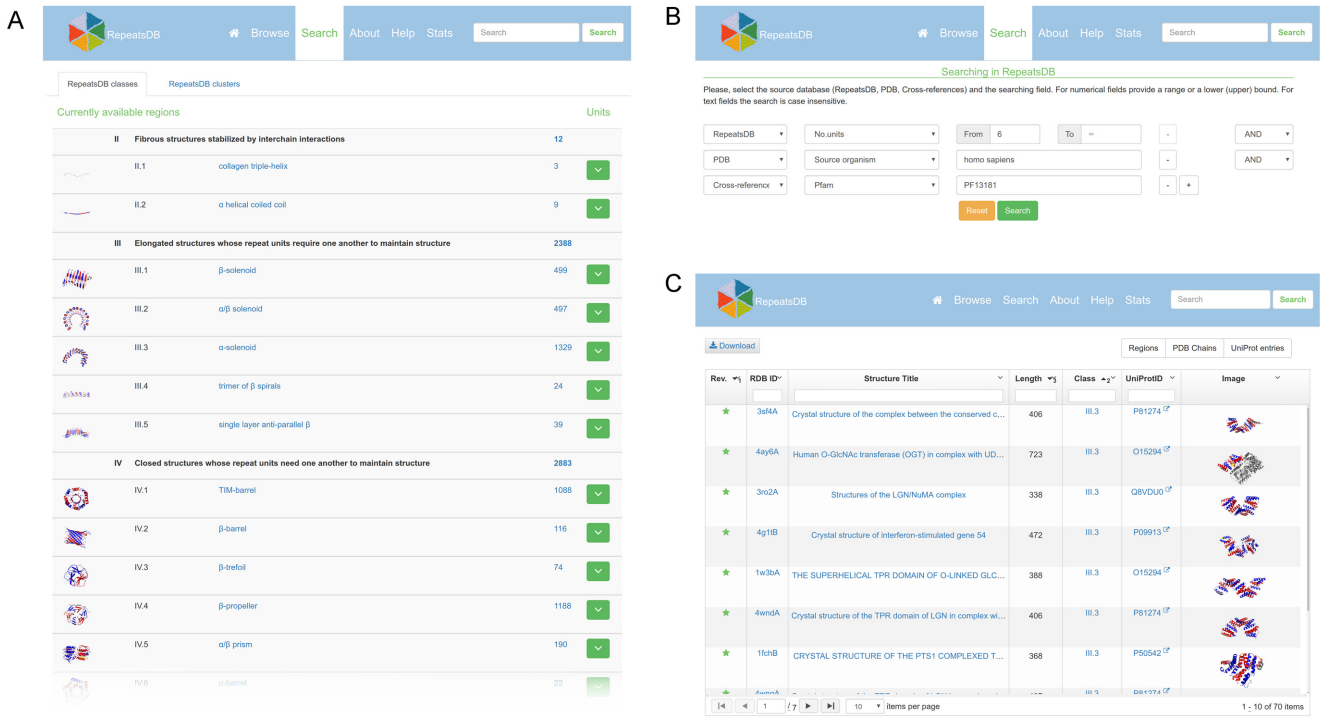

Figure 1. Retrieving RepeatsDB data. RepeatsDB data can be retrieved in three different ways. (**A**) The 'Browse' page provides the entry point for both the structural hierarchy and sequence clusters. (**B**) The 'Search' page allows the user to perform advanced queries against a range of RepeatsDB-specific and third-party search fields. The input can be simple text or numeric (single value or range) according to the field type and multiple queries can be combined by boolean operators (AND, OR, NOT). Both the 'Browse' and 'Search' pages redirect to the results page (**C**). This page provides a table with the list of retrieved entries and can be further filtered (and sorted) through column header fields. Results can be displayed by PDB chain (default), region or UniProt.

collagens and α-helical coiled coils, with various arrangements described in (22). Class III contains the most typical repeat examples, elongated structures where repetitive units require one another to maintain folding, e.g. β-, α/β- and α-solenoids. Class IV includes all closed repeat structures. Widespread across all types of organisms, this class includes the TIM-barrel and β-propeller subclasses. Both class III and IV contain units of length between 10 and 50 residues. Class V has unit lengths >40 residues and groups 'beads on a string' repeats, whose repeat units are large enough to fold independently.

All subclasses are characterized by a strong structural conservation in repeat units frequently not clearly reflected in sequence. This is the main reason why domain sequence databases such as Pfam (23) and SMART (24) fail to detect a large number of repeats (25). Indeed, as most of the largest clusters of human sequence regions not covered by Pfam were found to be repeated (25,26). RepeatsDB was developed to fill this gap and provides the community with a high-quality resource of reliable datasets of repeat structures for various purposes. The first and most obvious goal achieved was to compare the structural classification of repeats with the sequence-based one (26). Other uses of RepeatsDB are the extraction of repeat datasets to discuss specific features (27,28), benchmarking both sequence- and structure-based repeat detection methods and discussing the role of proteins with repeats (17,29–33). The high-quality manually annotated set of RepeatsDB units (Structural Repeat Unit Library, SRUL) exploits ReUPred (34) to predict repeat unit position and classification. RepeatsDB 2.0 includes new annotations, an improved classification and completely redesigned web server and interface, to guarantee availability of data and better user experience in terms of database usability and look-and-feel.

DATABASE DESCRIPTION

RepeatsDB 2.0 data have been completely regenerated taking advantage of the new ReUPred predictor (34) for automatic detection of tandem repeat units. In the new database version all entries are annotated at the unit level, i.e. providing start and end position for each repeated segment, and classified at the subclass level. Compared to the old version, unit annotations have grown by more than an order of magnitude. A detailed description of the RepeatsDB annotation pipeline follows.

Data curation

The initial dataset for RepeatsDB is the entire PDB (20). Repeat candidates are extracted with RAPHAEL (19) and processed with ReUPred (34) to confirm the presence of repeat regions and provide detailed unit information. ReUPred is a predictor able to identify the position of repeated fragments by performing iterative structural alignments against a manually refined library of representative units. ReUPred is also able to assign the class and subclass by transferring this information from the unit library. The

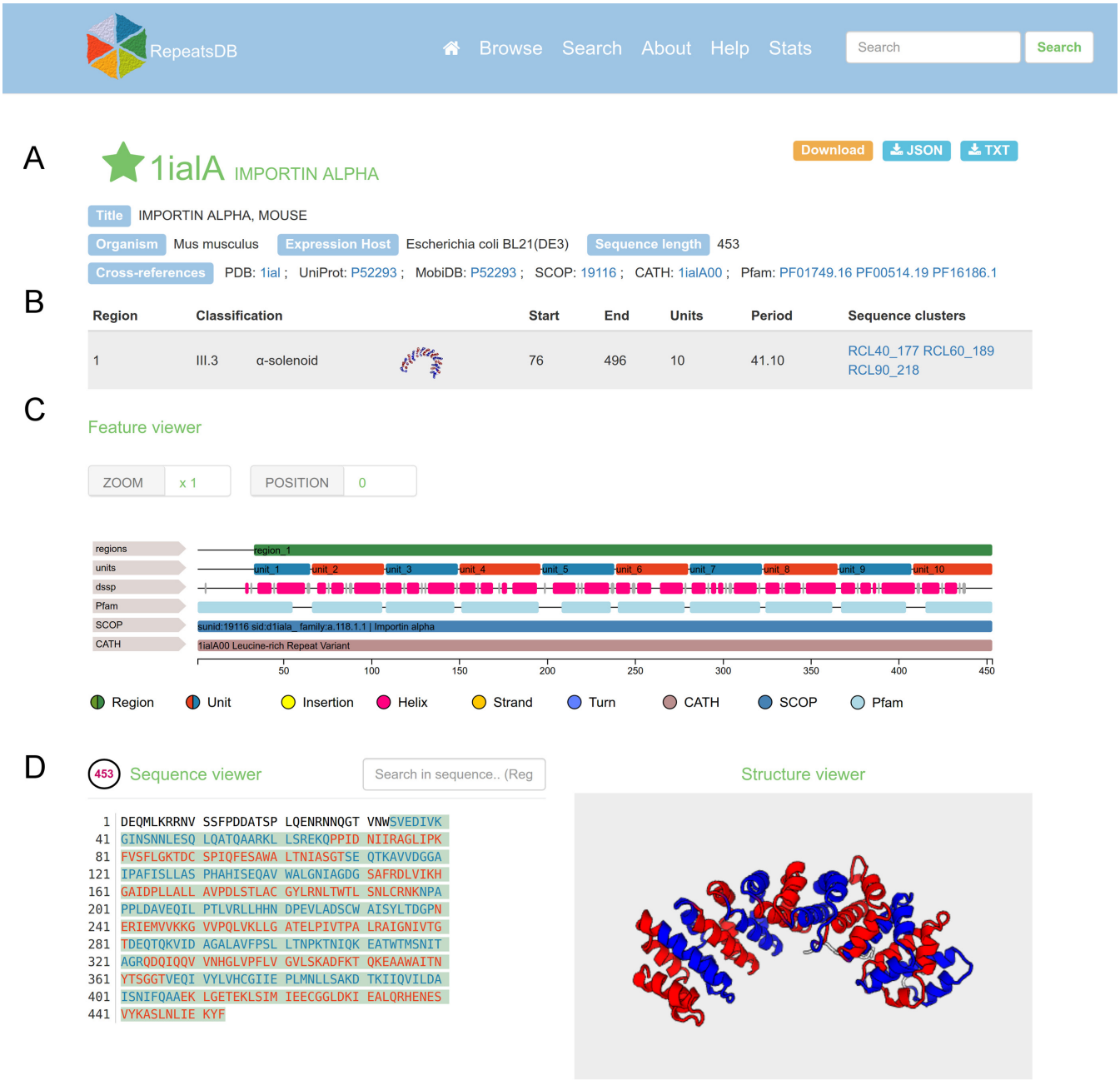

Figure 2. Screenshot of RepeatsDB sample entry page for PDB code 1ialA. The top part of the page (**A**) reports structure information from the PDB and cross-references to third-party databases including UniProt, MobiDB, SCOP, CATH and Pfam (when available). RepeatsDB annotations are available for download both in text and JSON formats on the top-right corner. (**B**) A table provides region details such as structural classification, start/end position, number of units, repeat period and cluster families. (**C**) The feature viewer summarizes available annotation for the PDB reference sequence, i.e. the SEQRES field in the PDB file. An overview of RepeatsDB information (regions, units and insertions) along with secondary structure (DSSP), Pfam, SCOP and CATH tracks (when available) are shown. (**D**) A detailed view of RepeatsDB annotations is highlighted in the sequence and PDB viewers.

final dataset available in RepeatsDB 2.0 is the result of an iterative process where the ReUPred library has been refined manually multiple times to resolve conflicts, improve its ability to generalize and include newly discovered subclasses. At the end of the process, an extensive validation and refinement of the predictions has been carried out by expert visual inspection. More than 60% of the entries have been reviewed and five new subclasses created, three for class IV (closed structures) and two for class V (beads on a string).

Implementation

RepeatsDB was designed as a multi-tier architecture, with three modules managing data storage, processing and presentation, respectively. Data are stored in a MongoDB database, and processed with Node.js. The server is accessi-

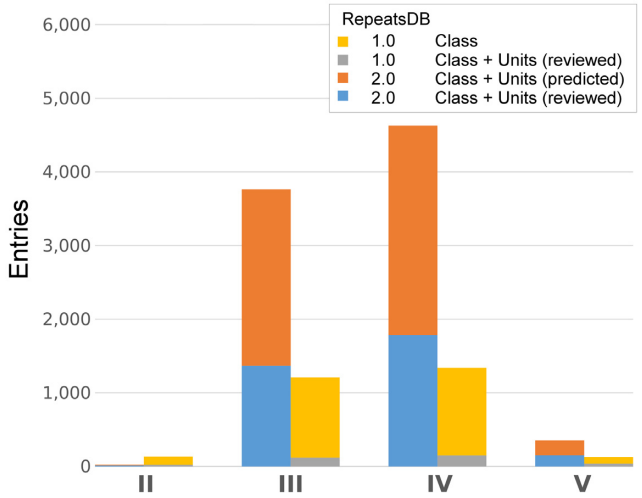

Figure 3. RepeatsDB growth. RepeatsDB 2.0 is compared to the previous release. Entries have unit and subclass annotation, with more than 60% manually reviewed (blue). For the old version, only a tiny fraction of entries have unit definition (cyan) and the rest is mostly annotated only at the class level (yellow).

ble through a web interface or programmatically exploiting a RESTful architecture. The web interface is designed using the Angular.js and Bootstrap frameworks. Dynamic and interactive elements of the entry page are developed using PV for structure visualization and Bio.js as sequence feature viewer, respectively. Both the database structure and Node.js server have been completely rewritten to improve efficiency and data reliability. Moreover, all data derived from third party resources have been processed and stored locally to prevent broken dependencies.

Innovations

Apart from the new annotation pipeline, several bugs have been fixed and many improvements have been introduced since the last RepeatsDB release. All positional annotations are now based on SIFTS (35), making them consistent with both PDB (20) and UniProt (36) references. The search engine has been completely redesigned. An intuitive search interface allows to perform complex queries using logical operators and guides the user through all possible searching fields. A new classification level has been added to include evolutionary relationships among different repeat regions. An all-vs.-all alignment of the repeat regions allowed to group them according to sequence similarity and to identify different repeat families. The new classification has been implemented as an independent layer on top of the existing structural features, and is available at three different identity thresholds (40%, 60% and 90%). The web interface allows to navigate entry clusters, providing an overview of the representative sequences inside each structural subclass.

DATABASE USAGE

The user interface presents an intuitive summary table providing direct access to all entries by structural class directly from the home page. For a finer search, the user can visit either the 'Browse' page providing subclass access or the 'Search' page for generating complex queries (Figure 1A and B). All entry points redirect to the same result page listing the retrieved proteins in a table (Figure 1C). The table can be further filtered by providing additional matching strings in the column headers. The 'Browse' page also provides direct access to sequence clusters, where entries are grouped by sequence similarity. The redesigned entry page (Figure 2) is much more informative compared to the previous RepeatsDB version, including several cross-links to third party resources. It also integrates several structural features useful for comparing CATH, SCOP, Pfam and DSSP annotations with RepeatsDB data. Regions, units and insertions are provided for all entries and correctly mapped both to UniProt and PDB reference (SEQRES field in the PDB file) sequences thanks to the SIFTS service. The correct mapping can strongly improve RepeatsDB impact since it is now very easy to link repeat data with other sequence features like mutations or post-translational modifications. Thanks to a RESTful architecture, all RepeatsDB data are accessible from external APIs and third party resources through HTTP URLs. Please refer to the 'Help' page of the website for details on using the RepeatsDB web services. Customized datasets can be downloaded in JSON or text format using the browse function or RESTful web services.

Statistics

RepeatsDB provides high quality annotation for ∼5400 entries. Figure 3 compares the current RepeatsDB content to the previous version. The chart shows the total number of entries belonging to each class. However, the new version provides unit definition and subclass classification for all entries where the old version annotated only a tiny fraction (327 entries, cyan bar). Moreover, in RepeatsDB 2.0 more than 60% of the entries have been manually reviewed by expert curators (blue segment). Further details such as the number of regions, units and genes are available from the 'Stats' page of the web site.

CONCLUSION AND FUTURE WORK

RepeatsDB was presented in 2014 with the goal to provide the community with a central resource for high-quality tandem repeat protein structure annotation. It has been cited in a number of different studies regarding repeat proteins, and has been used to extract datasets for repeat proteins analysis and to test algorithms for repeat proteins annotation. The detailed annotation of entries performed by RepeatsDB curators has allowed us to build of a high quality Structure Repeat Unit Library (SRUL). This library was exploited by the ReUPred algorithm (34) as a gold standard to define unit position in new entries in an iterative process.

The new release of RepeatsDB includes a new annotation pipeline, combining the RAPHAEL algorithm for repeat detection (19) and ReUPred for annotation (34), producing extensive annotation for all entries. The pipeline is fully automated and allows the easy regular update of the database. The iterative execution of the pipeline already demonstrated its efficacy both because it identified a large

number of new entries, and because new subclasses were identified and added to the structural classification scheme. RepeatsDB will benefit from regular updates, which will steadily increase the number of available annotations. Future work will concentrate on exploiting the repeat unit definitions to create profiles for use in detecting repeats from sequence for genome-scale analysis (37).

ACKNOWLEDGEMENTS

The authors are grateful to Francesco Tabaro for help with the web interface. We also acknowledge ELIXIR-IIB (elixir-italy.org), the Italian Node of the European ELIXIR infrastructure (elixir-europe.org), for supporting the development and maintenance of RepeatsDB.

FUNDING

COST Action BM1405 NGP-net. Funding for open access charge: COST BM1405.
Conflict of interest statement. None declared.

REFERENCES

1. Andrade,M.A., Perez-Iratxeta,C. and Ponting,C.P. (2001) Protein repeats: structures, functions, and evolution. *J. Struct. Biol.*, **134**, 117–131.
2. Kojic,S., Radojkovic,D. and Faulkner,G. (2011) Muscle ankyrin repeat proteins: their role in striated muscle function in health and disease. *Crit. Rev. Clin. Lab. Sci.*, **48**, 269–294.
3. Jha,S. and Ting,J.P.-Y. (2009) Inflammasome-associated nucleotide-binding domain, leucine-rich repeat proteins and inflammatory diseases. *J. Immunol. Baltim. Md*, **183**, 7623–7629.
4. Madsen,B.E., Villesen,P. and Wiuf,C. (2008) Short tandem repeats in human exons: a target for disease mutations. *BMC Genomics*, **9**, 410.
5. Orr,H.T. and Zoghbi,H.Y. (2007) Trinucleotide repeat disorders. *Annu. Rev. Neurosci.*, **30**, 575–621.
6. Liggett,W.H. and Sidransky,D. (1998) Role of the p16 tumor suppressor gene in cancer. *J. Clin. Oncol. Off. J. Am. Soc. Clin. Oncol.*, **16**, 1197–1206.
7. Cerveny,L., Straskova,A., Dankova,V., Hartlova,A., Ceckova,M., Staud,F. and Stulik,J. (2013) Tetratricopeptide repeat motifs in the world of bacterial pathogens: role in virulence mechanisms. *Infect. Immun.*, **81**, 629–635.
8. Weidle,U.H., Auer,J., Brinkmann,U., Georges,G. and Tiefenthaler,G. (2013) The emerging role of new protein scaffold-based agents for treatment of cancer. *Cancer Genomics Proteomics*, **10**, 155–168.
9. Vlassi,M., Brauns,K. and Andrade-Navarro,M.A. (2013) Short tandem repeats in the inhibitory domain of the mineralocorticoid receptor: prediction of a β-solenoid structure. *BMC Struct. Biol.*, **13**, 17.
10. Aziz,M.F., Caetano-Anollés,K. and Caetano-Anollés,G. (2016) The early history and emergence of molecular functions and modular scale-free network behavior. *Sci. Rep.*, **6**, 25058.
11. Alva,V., Söding,J. and Lupas,A.N. (2015) A vocabulary of ancient peptides at the origin of folded proteins. *eLife*, **4**, e09410.
12. Goncearenco,A. and Berezovsky,I.N. (2015) Protein function from its emergence to diversity in contemporary proteins. *Phys. Biol.*, **12**, 45002.
13. Jorda,J., Xue,B., Uversky,V.N. and Kajava,A.V. (2010) Protein tandem repeats - the more perfect, the less structured. *FEBS J.*, **277**, 2673–2682.
14. Abraham,A.-L., Rocha,E.P.C. and Pothier,J. (2008) Swelfe: a detector of internal repeats in sequences and structures. *Bioinformatics*, **24**, 1536–1537.
15. Parra,R.G., Espada,R., Sánchez,I.E., Sippl,M.J. and Ferreiro,D.U. (2013) Detecting repetitions and periodicities in proteins by tiling the structural space. *J. Phys. Chem. B*, doi:10.1021/jp402105j.
16. Hrabe,T. and Godzik,A. (2014) ConSole: using modularity of Contact maps to locate Solenoid domains in protein structures. *BMC Bioinformatics*, **15**, 119.
17. Do Viet,P., Roche,D.B. and Kajava,A.V. (2015) TAPO: A combined method for the identification of tandem repeats in protein structures. *FEBS Lett.*, **589**, 2611–2619.
18. Di Domenico,T., Potenza,E., Walsh,I., Parra,R.G., Giollo,M., Minervini,G., Piovesan,D., Ihsan,A., Ferrari,C., Kajava,A.V. *et al.* (2014) RepeatsDB: a database of tandem repeat protein structures. *Nucleic Acids Res.*, **42**, D352–D357.
19. Walsh,I., Sirocco,F.G., Minervini,G., Di Domenico,T., Ferrari,C. and Tosatto,S.C.E. (2012) RAPHAEL: recognition, periodicity and insertion assignment of solenoid protein structures. *Bioinformatics*, **28**, 3257–3264.
20. Velankar,S., van Ginkel,G., Alhroub,Y., Battle,G.M., Berrisford,J.M., Conroy,M.J., Dana,J.M., Gore,S.P., Gutmanas,A., Haslam,P. *et al.* (2016) PDBe: improved accessibility of macromolecular structure data from PDB and EMDB. *Nucleic Acids Res.*, **44**, D385–D395.
21. Kajava,A.V. (2012) Tandem repeats in proteins: from sequence to structure. *J. Struct. Biol.*, **179**, 279–288.
22. Moutevelis,E. and Woolfson,D.N. (2009) A periodic table of coiled-coil protein structures. *J. Mol. Biol.*, **385**, 726–732.
23. 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.
24. Letunic,I., Doerks,T. and Bork,P. (2015) SMART: recent updates, new developments and status in 2015. *Nucleic Acids Res.*, **43**, D257–D260.
25. Mistry,J., Coggill,P., Eberhardt,R.Y., Deiana,A., Giansanti,A., Finn,R.D., Bateman,A. and Punta,M. (2013) The challenge of increasing Pfam coverage of the human proteome. *Database*, **2013**, bat023.
26. Paladin,L. and Tosatto,S.C.E. (2015) Comparison of protein repeat classifications based on structure and sequence families. *Biochem. Soc. Trans.*, **43**, 832–837.
27. Brunette,T.J., Parmeggiani,F., Huang,P.-S., Bhabha,G., Ekiert,D.C., Tsutakawa,S.E., Hura,G.L., Tainer,J.A. and Baker,D. (2015) Exploring the repeat protein universe through computational protein design. *Nature*, **528**, 580–584.
28. Espada,R., Parra,R.G., Mora,T., Walczak,A.M. and Ferreiro,D.U. (2015) Capturing coevolutionary signals inrepeat proteins. *BMC Bioinformatics*, **16**, 207.
29. Turjanski,P., Parra,R.G., Espada,R., Becher,V. and Ferreiro,D.U. (2016) Protein repeats from first principles. *Sci. Rep.*, **6**, 23959.
30. Glantz,S.T., Carpenter,E.J., Melkonian,M., Gardner,K.H., Boyden,E.S., Wong,G.K.-S. and Chow,B.Y. (2016) Functional and topological diversity of LOV domain photoreceptors. *Proc. Natl. Acad. Sci. U.S.A.*, **113**, E1442–E1451.
31. Sharma,M. and Pandey,G.K. (2015) Expansion and function of repeat domain proteins during stress and development in plants. *Front. Plant Sci.*, **6**, 1218.
32. Parra,R.G., Espada,R., Verstraete,N. and Ferreiro,D.U. (2015) Structural and energetic characterization of the ankyrin repeat protein family. *PLoS Comput. Biol.*, **11**, e1004659.
33. Pellegrini,M. (2015) Tandem repeats in proteins: prediction algorithms and biological role. *Front. Bioeng. Biotechnol.*, **3**, 143.
34. Hirsh,L., Piovesan,D., Paladin,L. and Tosatto,S.C.E. (2016) Identification of repetitive units in protein structures with ReUPred. *Amino Acids*, **48**, 1391–1400.
35. Velankar,S., Dana,J.M., Jacobsen,J., van Ginkel,G., Gane,P.J., Luo,J., Oldfield,T.J., O'Donovan,C., Martin,M.-J. and Kleywegt,G.J. (2013) SIFTS: Structure integration with function, taxonomy and sequences resource. *Nucleic Acids Res.*, **41**, D483–D489.
36. UniProt: a hub for protein information (2015) *Nucleic Acids Res.*, **43**, D204–D212.
37. Mitchell,A., Chang,H.-Y., Daugherty,L., Fraser,M., Hunter,S., Lopez,R., McAnulla,C., McMenamin,C., Nuka,G., Pesseat,S. *et al.* (2015) The InterPro protein families database: the classification resource after 15 years. *Nucleic Acids Res.*, **43**, D213–D221.