W402–W407 *Nucleic Acids Research, 2018, Vol. 46, Web Server issue* *Published online 9 May 2018*
doi: 10.1093/nar/gky360

RepeatsDB-lite: a web server for unit annotation of tandem repeat proteins

Layla Hirsh[1,2,†], Lisanna Paladin[1,†], Damiano Piovesan[1] and Silvio C. E. Tosatto[1,3,*]

[1]Dept. of Biomedical Sciences, University of Padua, Padua, Italy, [2]Dept. of Engineering, Pontificia Universidad Católica del Perú, Lima, Perú and [3]CNR Institute of Neurosciences, Padua, Italy

Received March 06, 2018; Revised April 13, 2018; Editorial Decision April 23, 2018; Accepted April 24, 2018

ABSTRACT

RepeatsDB-lite (http://protein.bio.unipd.it/repeatsdb-lite) is a web server for the prediction of repetitive structural elements and units in tandem repeat (TR) proteins. TRs are a widespread but poorly annotated class of non-globular proteins carrying heterogeneous functions. RepeatsDB-lite extends the prediction to all TR types and strongly improves the performance both in terms of computational time and accuracy over previous methods, with precision above 95% for solenoid structures. The algorithm exploits an improved TR unit library derived from the RepeatsDB database to perform an iterative structural search and assignment. The web interface provides tools for analyzing the evolutionary relationships between units and manually refine the prediction by changing unit positions and protein classification. An all-against-all structure-based sequence similarity matrix is calculated and visualized in real-time for every user edit. Reviewed predictions can be submitted to RepeatsDB for review and inclusion.

INTRODUCTION

Tandem repeats (TR) in proteins are ubiquitous in genomes and have been demonstrated to be of fundamental importance in many biological processes (1). They have several unique functions related to the development of organism complexity (2). TR proteins are characterized by a modular structure stabilized by a pattern of local interactions (3), which can be arranged in a wide variety of shapes providing functional diversity. Structural TR modules, called units (4), correspond to repeated segments in the sequence which can be loosely conserved both at the DNA and amino acid level. Structural and functional properties of TR proteins are conserved even in presence of high divergence among subsequences and difficult to detect by sequence-based methods (5). TR proteins are classified based on structural topology (4) as elongated, closed or 'beads on a string' (3). A finer classification is possible considering the type and length of the composing units (3). Subtle differences in the structural conformation of the units give rise to large differences in shape and structural properties of the whole protein, including curvature and twist (6). For examples, families of closed structures are likely to conserve the number of units due to spatial constraints and are quite rigid. Elongated structures like solenoids are in general more flexible and the secondary structure and shape of the units provides a way to fine tune global properties (7). Protein function is also strongly related to unit types and single units are often recognized as functional determinants of protein families. For example, Ankyrin proteins are associated to binding-mediated inhibition of proteins, e.g. in the cyclin-dependent kinase inhibitor p16. Sequence profiles of TRs in Pfam (8) have been shown to broadly correspond to the structure-based TR classification (9). The problem of classifying TR proteins is far from being solved and in recent year a number of new families has been discovered. For example, β-hairpins has been identified in seven different folds including both closed and open structures (10). The largest collection of TR proteins detected by structural features is provided by the RepeatsDB database (11). It relies on computational approaches and expert manual curation to detect TR in the Protein Data Bank (PDB) structures (12). RepeatsDB annotates the unit position and provides a hierarchical classification of TR protein families at the level of classes and sub-classes (4). TR regions are classified as linearly arrayed structures, closed or 'beads on a string' of small domains (3).

At the time of writing RepeatsDB contains >6000 annotated entries, ∼60% of which are manually reviewed (11). The number of RepeatsDB entries is continuously growing thanks to regular updates, which require a continuous effort in the manual curation. It relies on structure-based methods providing the starting point for manual refinement based on visual analysis in a molecular viewer. Structure based methods detect TR periodicity by implementing different strategies, such as the identification of 3D symmetries or the analysis of regularity of structural features. In general,

*To whom correspondence should be addressed. Tel: +39 0498276269; Fax: +39 0498276260; Email: silvio.tosatto@unipd.it
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

these methods are designed to discriminate between repeat and non-repeat structures. In some cases they are able to annotate TR proteins with the position of units, e.g. ConSole and TAPO. ConSole (13) exploits the modularity of protein contact maps while TAPO (14) uses the periodicities of atomic coordinates and other types of structural representation. Our method ReUPred (15) identifies structural units by comparing the protein structure against a manually curated library of TR units. The TR unit library is designed to represent the conformational space and diversity of bona fide repeat units. It allows not only to identify the modular substructures in a TR, but also to classify them.

Here, we introduce RepeatsDB-lite, a web server designed for the detection and classification of TR proteins and the detection of repeated structural modules from PDB files. RepeatsDB-lite extends the ReUPred algorithm to all TR types and strongly improves the performance both in terms of computational time and accuracy. The web-based graphical user interface allows a complete visualization of the predictor results. Moreover, the server represents a platform to harness community annotation efforts, which have been proven to be effective in RepeatsDB experience.

MATERIALS AND METHODS

RepeatsDB-lite is a web server designed for the prediction, visualization and analysis of repeated regions in protein structures. It is based on an improved ReUPred algorithm (15) using several checks to minimize errors in the unit detection step and speeding up the calculation. A refactoring of the TR unit library allows it to cover all RepeatsDB classes. Its ability of predicting unit position is evaluated against all manually curated RepeatsDB entries (11). A comparison with existing methods is provided for a limited set of solenoid examples for which predictions are available (15). The web interface is designed to provide a complete visualization of the data including structural and sequence alignments of the predicted units. In addition, it includes an intuitive form to manually refine the annotation and visualize the effect on the unit alignments on the fly.

RepeatsDB-lite algorithm

RepeatsDB-lite is the evolution of the ReUPred method. It uses an iterative structural search against a library of TR units to find repetitive elements in protein structures. The inputs are a target structure and the TR unit library, which represents the conformational space and diversity of bona fide repeat units (see below). The algorithm exploits the library by aligning it against the target structure, using a divide and conquer approach. Once the best unit is identified by structural similarity with the library (called Master unit), the unit is fixed and the algorithm forks (divides). Only alignments satisfying the similarity criteria described in Table 1 are considered valid. Coverage, RMSD and TM-score thresholds were calculated from a similarity network analysis and guarantee separation between subclasses. Two new input structures are created, corresponding to the N- and C-terminal flanking fragments of the predicted unit and two new cycles (structural searches) are performed. The structural search on these two fragments is performed using a new ad hoc library created on the fly, populated by the Master unit and all newly predicted units which are included for search in the following cycles. With this approach, the repeat region is expanded until the new input fragments are too short, so the protein is processed, or the structural search does not provide any new valid alignment. The predicted units are then collected and evaluated together (conquer). In this phase, fragments included in the region but deviating from the classical unit structure are annotated as insertions. At the end, if the region has fewer than three units, the next potential unit from the library is used as Master and the entire iterative part is repeated from the beginning for up to four increasingly relaxed iterations. Compared to ReUPred, the new algorithm discards structural alignments where unit boundaries break (fall inside) secondary structure elements and is now able to detect multiple regions inside the same chain. For α-solenoids and β-hairpins (BHR) it also provides a finer classification that describes the unit conformational 'type' or 'fold', often corresponding to the protein family. The average execution time for a single chain is some minutes but varies depending on the class of the master unit.

TR unit library

The RepeatsDB-lite TR unit library represents all known repeat conformations, including elongated, closed and beads-on-a-string repeats, for a total of 20 subclasses. To increase predictor speed the unit library has been structured hierarchically in three layers. The search for the Master unit starts from the reduced library and then propagates to other layers considering only related units, i.e. belonging to the same cluster as the previous layer. The bottom layer of the unit library is built considering all units of manually curated RepeatsDB entries. A strong reduction is performed by excluding those with insertions (3,401 units) and with missing (non-crystallized) residues (530 units). Units diverging from the subclass average length and redundant units at 70% sequence identity, calculated with CD-HIT (16), are also discarded. At the end, the bottom layer counts a total of 2591 TR units. The other two layers are generated by reducing the structural similarity. Units are clustered at 0.5 TM-score and 80% overlap (coverage) in the middle layer (1160 units) and at 0.3 TM-score and 80% overlap for the SRUL core (top layer, 536 units).

Unit similarity

The RepeatsDB-lite software includes additional modules to analyze TR unit predictions. The first is a multiple structure alignment of the units calculated with Mustang (17) useful to highlight overall unit conformation, insertions, diverging units and prediction errors. Another output is a matrix representing the structural similarity between unit pairs. It is calculated by performing an all-against-all pairwise structure alignment with TM-align (18). The units in the matrix are reported from N- to C-terminus and cells are colored based on the observed sequence similarity calculated upon structural alignment and normalized by the length of the shortest unit. From the matrix it is possible to identify patterns of similarity useful to trace the evolutionary history of duplication events.

Table 1. Validation rules for RepeatsDB-lite predictions. The criteria used to include structural alignments in RepeatsDB-lite predictions are shown per repeat class. Coverage is the fraction of residues covered compared to the reference structure. RMSD is the root mean square deviation. The TM-score method is used to calculate the RMSD and TM-score values.

Alignment type	Class	Min coverage	Max RMSD	Min TM-score
Repeat unit library	III—Elongated	0.8	2.3	0.35
	IV—Toroid	0.6	2.8	0.4
	V—Beads on a string	0.6	3.5	0.4
Intra-protein	III—Elongated	—	2	0.25
	IV—Toroid	—	4	0.28
	V—Beads on a string	—	4	0.28

Web server implementation

The RepeatsDB-lite web server is implemented using the REST (Representational State Transfer) architecture, allowing access from a web-based user interface as well as programmatically through external APIs or third-party web services exploiting the Node.js functionality. The web interface has been developed using the Angular.js framework and Bootstrap CSS style sheets. Dynamic and interactive elements of the output page are developed using PV (https://biasmv.github.io/pv/) for structure visualization and BioJS (https://biojs.net/) for sequence alignments.

Benchmarking unit prediction

In order to characterize TR proteins it is necessary to identify the position of the repetitive structural elements (units). Assuming TR units are structurally similar inside the same protein, the problem would be reduced to the identification of the unit phase and length. Since in reality units are not homogeneous but often include insertions and structural variation, the evaluation has to be performed unit by unit. We considered manually curated RepeatsDB (version 2017.10.25) entries as source of real unit annotation for benchmarking. TAPO and Console (13,14) were also compared on the solenoids class. The dataset is the same used in the ReUPred paper (same proteins) with updated unit annotation according to the latest RepeatsDB release. Unit prediction performance is measured adopting a strategy similar to the ReUPred paper (15). To obtain a fairer evaluation and assess the effect of incomplete data, RepeatsDB-lite was also benchmarked removing units in the library with over 40%, 60% or 80% sequence identity with dataset proteins. Each reference unit is paired with the predicted unit with maximum symmetric coverage (if any). True positives (TP) are matching residues, false positives (FP) are all predicted unit residues outside reference units, false negatives (FN) are reference unit residues not overlapping with any matching unit and true negatives (TN) are all residues correctly predicted as not repeated. Insertion residues in the reference are masked, i.e. not considered for the calculation of the confusion matrix. Predicted insertions are considered negative predictions and overwrite overlapping unit predictions. When the predictor does not identify any unit or the returned file is empty it is evaluated as a fully negative prediction. When the reference protein contains multiple TR regions the entire sequence is split along the middle point between regions. This is necessary to distribute negative residues equally between regions and to perform region and class based statistics accurately. Sensitivity (Sn = TP/(TP + FN)), specificity (Sp = TN/(TN + FP)), precision (Pr = TP/(TP + FP)), balanced accuracy (Acc = (Sn + Sp)/2) and F-measure (F = 2 * Pr * Sn/(Pr + Sn)) are calculated.

SERVER DESCRIPTION

RepeatsDB-lite takes a PDB structure in input and predicts TR units and the repeat classification along the RepeatsDB schema (class, subclass, type and fold). The server accepts either a PDB identifier (ID) or file. By default, the predictor considers only the first PDB chain. Alternatively, the user can specify the chain ID or an 'all chains' mode. Submitted jobs can be retrieved using the search box or bookmarking the result page URL. The RepeatsDB-lite output page features an intuitive visualization of predicted TR regions and units. In addition, another page allows the user to modify the prediction and visualize the effect on the unit alignments on the fly. The reviewed prediction can optionally be submitted for review and inclusion in RepeatsDB.

Output page

The different visualizations contained in the RepeatsDB-lite output page (Figure 1, panel A) are designed to guide the analysis of the repeat structure. The page header (Figure 1A, top) provides general information such as the name of the input PDB (or file), processed chains and session identifier. Multiple chains are visualized in different tabs (Figure 1A, middle). When multiple regions (groups of units) are identified in the same chain, they are visualized in the same page separated in different blocks. For each chain, regions and units are visualized in a structure and sequence viewer. For each region (Figure 1A, bottom) the multiple structure alignment of the units and resulting sequence and secondary structure alignments are visualized along with the similarity matrix (see unit similarity paragraph) representing sequence similarity based on an all-against-all structure alignment. The PDB input, the predictor output, log and mapping files (between position along the SEQRES and PDB indices of each residue) are available for download. The manual refinement of the unit in the single chain can be accessed through the 'Edit annotation' button in the corresponding chain tab.

Edit page

RepeatsDB-lite includes a page for the manual refinement of unit annotations (Figure 1, panel B). The form fields (Figure 1B, left) allow the curator to add/delete regions,

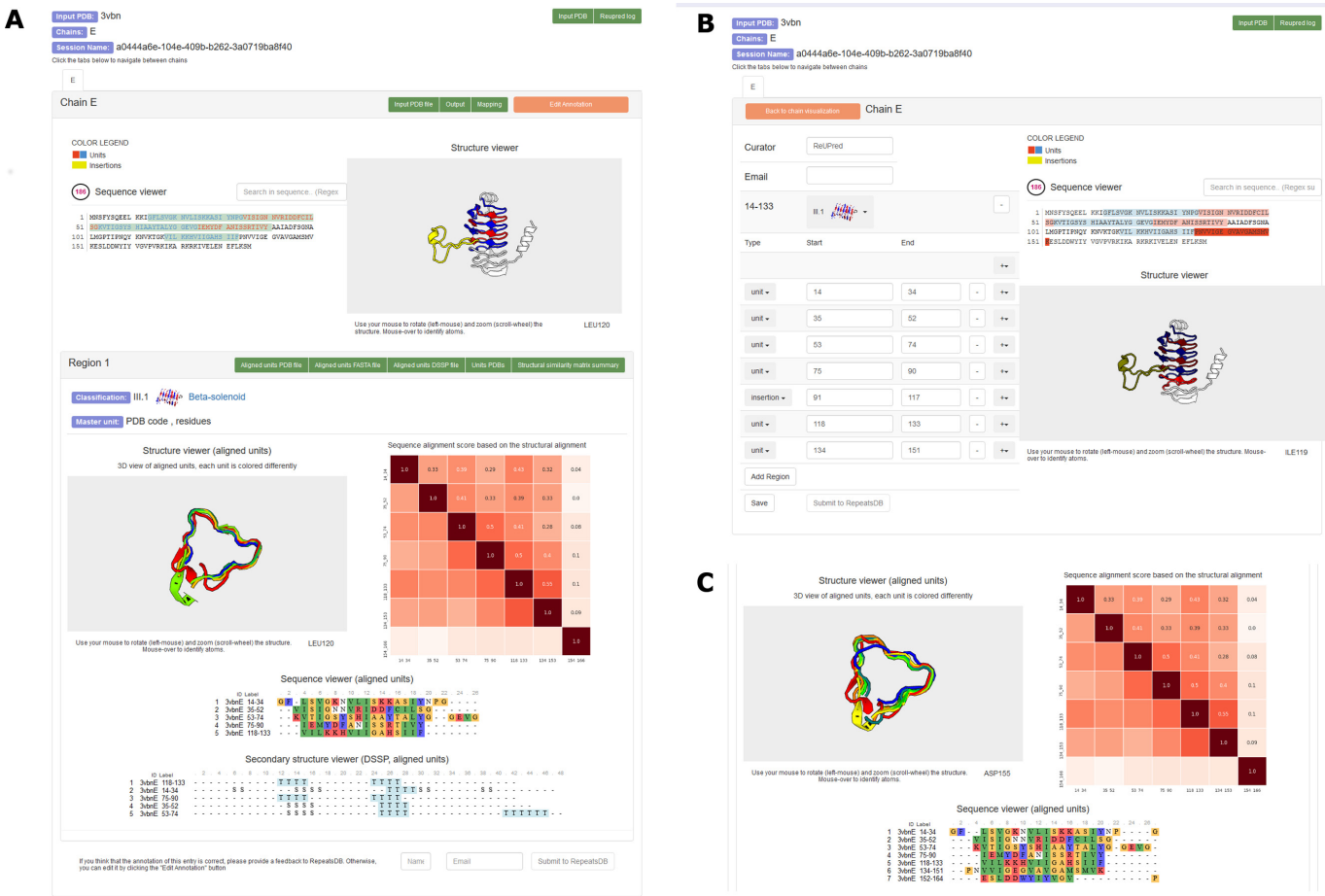

Figure 1. RepeatsDB-lite pages. (**A**) Result page. The header provides summarizing information about the job (PDB code: 3vbn). The tabs below allow the navigation between chain predictions. Each chain tab shows some general information about the chain and a specific card for each region. Download buttons allow the retrieval of text file results, while sequence, structure and alignment viewer guide data visualization. The unit sequence similarity matrix shows the relationship between units in the region. The orange button redirects to the form for annotation editing. (**B**) Annotation editing page. The user can modify region classification and the unit start end position. Changes are reflected in the viewers on the right. Reviewed annotation can be submitted directly to RepeatsDB maintainers to be included in the database. (**C**) Results after resubmission. By saving the repeat annotation edits, the user is redirected to RepeatsDB output page where he is provided a detailed visualization of new results to evaluate the annotation quality.

change classification and modify unit annotation. On the right side of the page (Figure 1B, right), a sequence and structure viewer react to the user edits, allowing a preliminary evaluation of the changes. Upon clicking the 'Submit' button, the user is redirected to the results page whose content is updated according to the provided new annotation. Finally, the 'Submit to RepeatsDB' button, available both in the edit page and in the output page, allows to submit the curated annotation to RepeatsDB for review and inclusion.

Usage example

The AntD *N*-acyltransferase from *Bacillus cereus* (PDB code: 3vbn) forms a ternary complex of three solenoid chains (19). Each chain folds into a left-handed β-helix of seven turns, interrupted by a loop and ending with an α-helix. The loop extends toward the flanking subunits and provides a binding platform for the ligands (Coenzyme A and dTDP). RepeatsDB-lite correctly identifies five β-solenoid elements and the 27 residue long insertion between the fourth and fifth unit (Figure 1, panel A). It is possible to appreciate the good phasing from the multiple structure alignment. In addition, the similarity matrix shows some darker cells close to the diagonal indicating how adjacent units are more similar to each other compared than distant ones. Even if shorter, two other units are missing in the RepeatsDB-lite output. The user can add them from the edit page and immediately see the results in the sequence and structure viewer (Figure 1, panel B). By clicking the 'Save' button the user is redirected to the result page and the similarity matrix is recalculated as well as the sequence and structural alignment (Figure 1, panel C). The latter in particular shows how the added unit diverge slightly from the perfect superimposition pattern of the previous multiple alignment including only the first five units (Figure 1C, bottom left). The similarity matrix, where two additional elements are added (Figure 1C, right), shows how the last unit in particular diverges significantly from the others. The different visualizations are designed to guide the user in the annotation refinement process. Users are encouraged to send the reviewed annotation to the RepeatsDB maintainers by clicking the corresponding button.

Table 2. Comparison with other methods. The regions column corresponds to the number of evaluated TR regions, i.e. for which a predictor provides an output, including fully negative predictions (zero units). Sensitivity (Sn), specificity (Sp), precision (Pr), balanced accuracy (Acc) and F-measure (*F*) values are in the range [0, 1]. Best values are in bold.

		Method	Regions	Sn	Sp	Pr	Acc	*F*
III.1	β-solenoid	TAPO	31	0.546	0.802	0.851	0.674	0.665
		ConSole	31	0.510	0.811	0.848	0.661	0.637
		RepeatsDB-lite 40	31	0.398	0.959	0.952	0.678	0.561
		RepeatsDB-lite 60	31	0.543	0.962	0.967	0.752	0.695
		RepeatsDB-lite 80	31	0.560	0.912	0.929	0.736	0.699
		RepeatsDB-lite	31	**0.598**	**0.953**	**0.963**	**0.776**	**0.738**
III.2	α/β solenoid	TAPO	18	0.692	0.699	0.925	0.696	0.792
		ConSole	18	0.644	0.834	0.954	0.739	0.769
		RepeatsDB-lite 40	18	0.558	0.912	0.967	0.735	0.707
		RepeatsDB-lite 60	18	0.790	0.851	0.961	0.820	0.867
		RepeatsDB-lite 80	18	0.788	0.847	0.960	0.818	0.866
		RepeatsDB-lite	18	**0.838**	**0.864**	**0.971**	**0.851**	**0.900**
III.3	α-solenoid	TAPO	38	0.665	0.577	0.916	0.621	0.771
		ConSole	38	0.552	**0.820**	**0.955**	0.686	0.700
		RepeatsDB-lite 40	38	0.747	0.561	0.914	0.654	0.822
		RepeatsDB-lite 60	38	0.859	0.595	0.930	0.727	0.893
		RepeatsDB-lite 80	38	0.839	0.668	0.946	0.754	0.889
		RepeatsDB-lite	38	**0.885**	0.684	0.951	**0.784**	**0.917**
III		TAPO	87	0.630	0.747	0.898	0.688	0.740
		ConSole	87	0.556	0.823	0.917	0.690	0.693
		RepeatsDB-lite 40	87	0.589	0.849	0.932	0.719	0.722
		RepeatsDB-lite 60	87	0.737	0.850	0.945	0.793	0.828
		RepeatsDB-lite 80	87	0.733	0.844	0.944	0.788	0.825
		RepeatsDB-lite	87	**0.778**	**0.855**	**0.950**	**0.816**	**0.855**

Table 3. RepeatsDB-lite performance against RepeatsDB reviewed entries. Columns headers have the same meaning of Table 2.

	Classification	Regions	Sn	Sp	Pr	Acc	*F*
II.1	Collagen triple-helix	3	0.000	0.000	0.000	0.000	0.000
II.2	α helical coiled coil	9	0.594	0.865	0.875	0.730	0.708
II	Fibrous repeats	12	0.545	0.865	0.875	0.705	0.672
III.1	β-Solenoid	325	0.561	0.926	0.911	0.743	0.694
III.2	α/β solenoid	350	0.797	0.907	0.984	0.852	0.881
III.3	α-Solenoid	888	0.784	0.628	0.943	0.706	0.856
III.4	β trefoil / β hairpins	76	0.583	0.960	0.968	0.772	0.728
III.5	Anti-parallel β layer / β hairpins	63	0.642	0.723	0.869	0.683	0.739
III	Elongated repeats	1702	0.750	0.819	0.948	0.785	0.838
IV.1	TIM-barrel	538	0.669	0.731	0.932	0.700	0.778
IV.2	β-Barrel / β hairpins	77	0.682	0.863	0.970	0.772	0.801
IV.3	β-Trefoil	24	0.449	0.754	0.731	0.602	0.556
IV.4	β-propeller	849	0.677	0.845	0.968	0.761	0.797
IV.5	α/β prism	185	0.782	0.956	0.997	0.869	0.876
IV.6	α-Barrel	18	0.419	0.931	0.917	0.675	0.576
IV.7	α/β barrel	5	0.986	0.000	0.995	0.493	0.991
IV.8	α/β propeller	117	0.591	0.836	0.933	0.713	0.723
IV.9	α/β trefoil	70	0.836	0.929	0.973	0.883	0.899
IV.10	Aligned prism	45	0.856	0.978	0.998	0.917	0.921
IV	Closed repeats	1928	0.685	0.826	0.961	0.755	0.800
V.1	α-Beads	13	0.758	0.652	0.980	0.705	0.855
V.2	β-Beads	42	0.813	0.779	0.975	0.796	0.887
V.3	α/β-beads	14	0.296	0.864	0.990	0.580	0.456
V.4	β sandwich beads	37	0.429	0.850	0.984	0.639	0.597
V.5	α/β sandwich beads	48	0.452	0.698	0.969	0.575	0.616
V	Beads on a string	154	0.537	0.759	0.975	0.648	0.692
All		3796	0.706	0.821	0.956	0.764	0.812

RepeatsDB-lite performance

RepeatsDB-lite is able to predict all types of TR proteins. In Table 2, a comparison with other methods is provided. The benchmark is the same used previously for ReUPred (15) but with updated unit annotations, i.e. considering reviewed information from the last RepeatsDB release. The dataset includes 87 solenoid regions from 84 proteins with 679 units for a total of 19 646 repeat and 5560 non-repeat residues. The region column corresponds to the evaluated regions. RepeatsDB-lite consistently reaches a precision above 95% and outperforms the other methods both considering balanced accuracy and F-measure (Table 2). ConSole has a better precision for α-solenoids at the cost of missing about half of the truly repeated residues (low sensitivity). Filtered versions of the RepeatsDB-lite unit library at 80, 60 and 40% sequence identity are benchmarked to assess the ef-

fects of redundancy with the test dataset. RepeatsDB-lite still shows a good accuracy even at 40% identity. In order to evaluate unit detection accuracy with a higher significance, RepeatsDB-lite was evaluated against all reviewed entries of RepeatsDB, for a total of 3666 proteins with 3835 TR regions and 29 113 units. The dataset contains 1 051 562 repeated and 193 338 non repeated residues (i.e. outside TR units). Insertion residues (29 403) are masked, i.e. not considered in the evaluation. Considering them as negatives does not affect the performance (data not shown). Results for the entire dataset and each subclass are reported in Table 3. RepetasDB-lite provides prediction for 3628 proteins, 136 of which contain multiple regions. α/β solenoid (III.2), α/β prism (IV.5), α/β trefoil (IV.9) and aligned prism (IV.10) are the best predicted subclasses, with a balanced accuracy over 0.8. In general, the majority of the examples come from class III and IV with the former having better sensitivity and the latter better specificity. RepeatsDB-lite fails when the unit length and structure diverge too much. Class IV has a larger unit structural variability that is remarkable also inside the same region. Another source of errors are those cases for which a single unit in the reference corresponds to multiple units in the prediction (or vice versa). Even when a unit perfectly matches multiple units in the counterpart, these cases are strongly penalized because the evaluation algorithm selects at most one match for each reference unit and counts non-overlapping residues as false negatives. Class V contains globular 'bead' domains and the size of the dataset is much smaller than the other two. In this case RepeatsDB-lite accuracy is lower because it identifies repetitions inside domains that are generally annotated as single units by curators. Class II includes single helix fibrous structures stabilized by inter-chain interactions and lack structural repetitions. As unit annotation is completely arbitrary and unrelated to structural properties, any evaluation can be considered meaningless. RepeatsDB-lite is also able to detect the structural classification of the TR region. In particular, it correctly detects the subclass for 77% of class III proteins and 80% of class IV (data not shown).

CONCLUSIONS

TR proteins are increasingly studied as evidence for new functions accumulate and new structures become available. The precise identification of the structural repeat modules allows to infer structural properties of the entire protein, its family and function. RepeatsDB-lite allows to identify units and classify the protein exploiting a structural similarity search and the information available in RepeatsDB. It outperforms existing methods and can be applied to all types of TR proteins. The web interface allows to visualize similarity relationships between TR units at both the sequence and structure level. The prediction can be manually refined by the user, visualizing the effects of the edits in real time. Annotations can be submitted to RepeatsDB for reviewed and we hope this will increase the amount of community-curated entries in the database. RepeatsDB-Lite can be seen as an example of gamification principles to engage a wider community towards database curation.

ACKNOWLEDGEMENTS

The authors are grateful to Andrey Kajava for insightful discussions.

FUNDING

University of Padua. Funding for open access charge: University of Padua.
Conflict of interest statement. None declared.

REFERENCES

1. Marcotte,E.M., Pellegrini,M., Yeates,T.O. and Eisenberg,D. (1999) A census of protein repeats. *J. Mol. Biol.*, **293**, 151–160.
2. Schaper,E., Gascuel,O. and Anisimova,M. (2014) Deep conservation of human protein tandem repeats within the eukaryotes. *Mol. Biol. Evol.*, **31**, 1132–1148.
3. Kajava,A.V. (2012) Tandem repeats in proteins: from sequence to structure. *J. Struct. Biol.*, **179**, 279–288.
4. Di Domenico,T., Potenza,E., Walsh,I., Gonzalo Parra,R., 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**, 1–6.
5. Schaper,E., Kajava,A.V., Hauser,A. and Anisimova,M. (2012) Repeat or not repeat?—statistical validation of tandem repeat prediction in genomic sequences. *Nucleic Acids Res.*, **40**, 10005–10017.
6. Kobe,B. and Kajava,A.V. (2000) When protein folding is simplified to protein coiling: the continuum of solenoid protein structures. *Trends Biochem. Sci.*, **25**, 509–515.
7. Kajava,A.V. (2002) What curves alpha-solenoids? Evidence for an alpha-helical toroid structure of Rpn1 and Rpn2 proteins of the 26 S proteasome. *J. Biol. Chem.*, **277**, 49791–49798.
8. 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.
9. 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.
10. Roche,D.B., Viet,P.D., Bakulina,A., Hirsh,L., Tosatto,S.C.E. and Kajava,A.V. (2018) Classification of β-hairpin repeat proteins. *J. Struct. Biol.*, **201**, 130–138.
11. Paladin,L., Hirsh,L., Piovesan,D., Andrade-Navarro,M.A., Kajava,A.V. and Tosatto,S.C.E. (2017) RepeatsDB 2.0: improved annotation, classification, search and visualization of repeat protein structures. *Nucleic Acids Res.*, **45**, D308–D312.
12. Burley,S.K., Berman,H.M., Kleywegt,G.J., Markley,J.L., Nakamura,H. and Velankar,S. (2017) Protein data bank (PDB): the single global macromolecular structure archive. *Methods Mol. Biol. Clifton NJ*, **1607**, 627–641.
13. Hrabe,T. and Godzik,A. (2014) ConSole: using modularity of contact maps to locate solenoid domains in protein structures. *BMC Bioinformatics*, **15**, 119.
14. 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.
15. 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.
16. Huang,Y., Niu,B., Gao,Y., Fu,L. and Li,W. (2010) CD-HIT suite: a web server for clustering and comparing biological sequences. *Bioinformatics*, **26**, 680–682.
17. Konagurthu,A.S., Whisstock,J.C., Stuckey,P.J. and Lesk,A.M. (2006) MUSTANG: a multiple structural alignment algorithm. *Proteins*, **64**, 559–574.
18. Zhang,Y. and Skolnick,J. (2005) TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res.*, **33**, 2302–2309.
19. Kubiak,R.L. and Holden,H.M. (2012) Structural studies of AntD: an N-Acyltransferase involved in the biosynthesis of D-anthrose. *Biochemistry (Mosc.)*, **51**, 867–878.