Amino Acids (2016) 48:1391–1400
DOI 10.1007/s00726-016-2187-2

ORIGINAL ARTICLE

Identification of repetitive units in protein structures with ReUPred

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

Received: 22 January 2016 / Accepted: 23 January 2016 / Published online: 22 February 2016

Abstract Over the last decade, numerous studies have demonstrated the fundamental importance of tandem repeat (TR) proteins in many biological processes. A plethora of new repeat structures have also been solved. The recently published RepeatsDB provides information on TR proteins. However, a detailed structural characterization of repetitive elements is largely missing, as repeat unit annotation is manually curated and currently covers only 3 % of the bona fide TR proteins. Repeat Protein Unit Predictor (ReUPred) is a novel method for the fast automatic prediction of repeat units and repeat classification using an extensive Structure Repeat Unit Library (SRUL) derived from RepeatsDB. ReUPred uses an iterative structural search against the SRUL to find repetitive units. On a test set of solenoid proteins, ReUPred is able to correctly detect 92 % of the proteins. Unlike previous methods, it is also able to correctly classify solenoid repeats in 89 % of cases. It also outperforms two recent state-of-the-art methods for the repeat unit identification problem. The accurate prediction of repeat units increases the number of annotated repeat units by an order of magnitude compared to the sequence-based Pfam classification. ReUPred is implemented in Python for Linux and freely available from the URL: http://protein.bio.unipd.it/reupred/.

Keywords Repeat protein · Structure prediction · Protein classification

Layla Hirsh and Damiano Piovesan Contributed equally.

✉ Silvio C. E. Tosatto
silvio.tosatto@unipd.it

[1] Department of Biomedical Sciences, University of Padua, Padua, Italy

[2] Department of Engineering, Pontificia Universidad Católica del Perú, Lima, Perú

[3] CNR Institute of Neuroscience, Padua, Italy

Introduction

Tandem repeat (TR) proteins characterized by a repetitive 3D structure have been exploited by nature in a myriad different cellular pathways and organisms (Marcotte et al. 1999b; Kobe and Kajava 2000; Kajava 2001, 2012). They are widely distributed in archeal, bacterial and eukaryotic proteomes and prevalent in complex organisms. An association was suggested between TR spread and the evolution of multicellularity (Marcotte et al. 1999a). Characterized by repetitions in their coding sequence, TRs are believed to have arisen from the duplication of short coding DNA segments (Andrade et al. 2001). These repetitions in sequence account for a peculiar modular fold architecture (Kajava 2012). Each structural module of this architecture is a "unit", the assembly of at least three of these building blocks forming a repeat "region" (Di Domenico et al. 2014). TR protein classification is based on repeat unit length (Kajava 2012), which can vary from one or two residues in crystallites (class I) to more than 50 residues in beads-on-a-string (class V), TR proteins built from the repetition of small globular domains (Kajava 2012). The middle ground comprises elongated (class III) and closed (class IV) structures, but it is dominated by the presence of a subtype of elongated structures, called solenoids (Kobe and Kajava 2000; Kajava 2001). Mainly due to stabilizing intra-unit short-range interactions, these proteins can be extended and refolded when subjected to a mechanical stretch force (Kim et al. 2010). In addition, they easily tolerate insertion of new units and possess an easily tunable horseshoe shape (Bazan and Kajava 2015). These

exceptional properties render them very efficient for protein–protein interactions (Andrade et al. 2001) and account for their widespread presence in cellular pathways. There has been an increasing interest in TR proteins and solenoids in particular, over the last few years, mainly due to their relevance in health (de Wit et al. 2011; Fournier et al. 2013) and for engineering applications (Grove et al. 2008; Höcker 2014; Brunette et al. 2015). TR domains have been exploited for the design of target-specific binders thanks to their favorable expression and stability properties (Binz et al. 2004; Varadamsetty et al. 2012), taking advantage of available data on natural repeat proteins to have control of binding surface and shape (Parmeggiani et al. 2008; Park et al. 2015). However, this class of proteins still largely belongs to the "dark matter" of the protein universe being characterized by non-canonical sequence–structure relationships. According to a study intended to assess the Pfam coverage of the human proteome, TR domains fall into the less characterized clusters of protein sequences (Mistry et al. 2013). Indeed, they evolve quickly while maintaining their fold, hampering detection by traditional methods for sequence analysis. The same holds for modeling and functional characterization, which usually relies on well-conserved sequence features. As a result, specialized methods were built for the identification of repeat proteins (Pellegrini 2015). Sequence-based strategies, based on homology search (Andrade et al. 2000) or domain assignment (Finn et al. 2014; Mitchell et al. 2015), mostly underestimate TRs due to the presence of highly degenerate repeat units (Mistry et al. 2013). Alternatively, methods requiring no prior knowledge for the detection of repeated substrings can be based on self-comparison (Heger and Holm 2000; Szklarczyk and Heringa 2004), clustering (Newman and Cooper 2007; Jorda and Kajava 2009) or hidden Markov models (Söding et al. 2006; Biegert and Soding 2008). Some others rely on complexity measurements (Pellegrini et al. 2012) or take advantage of meta searches to combine outputs from different sources (Gruber et al. 2005; Schaper et al. 2015). Methods recognizing TR proteins based on the modularity of their 3D structure have also been developed (Abraham et al. 2008; Sabarinathan et al. 2010; Walsh et al. 2012; Hrabe and Godzik 2014; Do Viet et al. 2015). To explore the relationship between repeat sequence and structure detection and classification, a comparison was performed between Pfam repeat families and RepeatsDB entries (Paladin and Tosatto 2015). RepeatsDB (Di Domenico et al. 2014) represents the state of the art for the annotation of tandem repeat structures. The database adds to the typical classification at a subclass level based on secondary and tertiary structure features. Existing methods for TR protein identification do not deal with the TR structures classification problem, which was based on manual assignment in RepeatsDB. RepeatsDB provides the start and end position of repetitive units only for a small subset ("detailed") which are manually annotated. Another group of proteins is provided only with the manual classification or by sequence similarity. The rest of the structures ("predicted") lack any classification and each represents the majority of the data.

The identification of single repeat units has so far been addressed by few automatic methods. ConSole (Hrabe and Godzik 2014) exploits the modularity of protein contact maps and TAPO (Do Viet et al. 2015), the periodicities of atomic coordinates and other types of structural representation. Both are available through a Web server interface that allows the user to evaluate one protein at a time. The automatic identification of units inside a TR protein structure allows to scale up this type of information. This newly available data could be a powerful tool to understand TR evolution and assess conservation at the sequence level. Collection of an "alphabet" of TR units can also be useful for protein engineering applications (Brunette et al. 2015). Here, we present a new Repeat Protein Unit Predictor (ReUPred) for the classification and identification of repetitive elements in TR proteins. It concentrates on the solenoid proteins, since it is one of the most abundant class of tandem repeat proteins in nature (Kajava 2001). The aim of ReUPred is to extend the detailed annotation for all classified proteins in RepeatsDB and possibly for all predicted repeats.

Methods

ReUPred is a predictor for the classification of tandem repeat proteins and identification of the composing repeat units. Its input is a target protein structure and a structural repeat unit library (SRUL). The output is a list of fragments corresponding to the predicted unit positions in the structure and the class assignment according to the RepeatsDB definition (Di Domenico et al. 2014). In this work, only solenoid repeat proteins (i.e., classes III.1 to III.3 in RepeatsDB) have been considered as they represent the most abundant class of repeat proteins in nature (Kajava 2001). The iterative algorithm decomposes the input structure using a template library. A divide and conquer strategy is used to improve both accuracy and speed, requiring on average ca. 2 min on a standard laptop. ReUPred was optimized by filtering SRUL, fine-tuning parameters to choose the best alignment and detect insertions between units as well as identifying separated repeat regions in the input protein. Each step is described in the following. ReUPred is implemented in Python for Linux. The source code is distributed under the GPL license and freely available from the URL: http://protein.bio.unipd.it/reupred/.

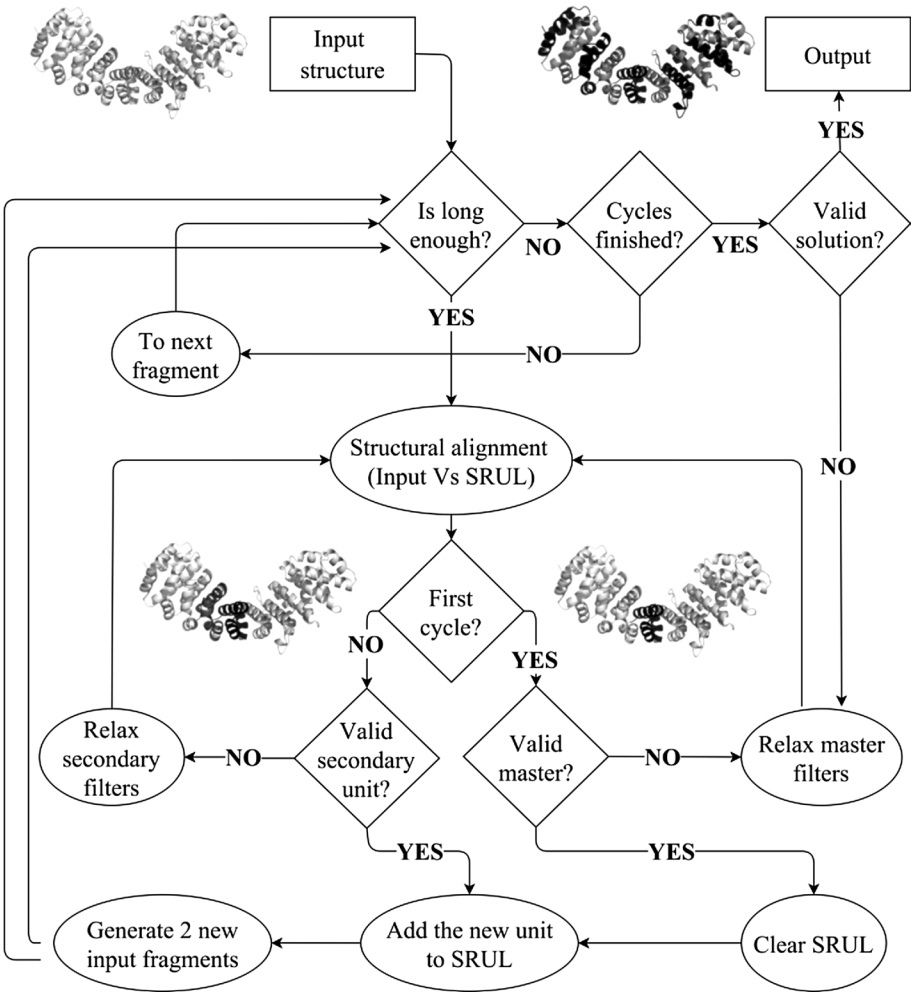

Fig. 1 Schematic description of the ReUPred algorithm. The input structure (PDB 1IQ1, chain C) is processed iteratively until a valid solution is provided and no new fragments (sub problems) are generated

ReUPred algorithm

The algorithm exploits the evolutionary history of tandem repeat proteins. Solenoid units have been demonstrated to evolve from a single representative unit to multiple copies through repeated duplications (Björklund et al. 2006) Units of a solenoid protein show a different degree of similarity, which is strongly correlated to the distance from the middle of the repeat region. This is consistent with the observation that units at the edges are more degenerated (Marcotte et al. 1999b). ReUPred exploits this knowledge and tries to mimic evolution. The objective is to predict adjacent units, i.e., to minimize the number of residues between predicted flanking units and obtain at least three repeated elements. This is important since in known RepeatsDB solenoid structures, insertions of non-repeat fragments are rare and mostly observed inside and not between units. See Fig. 1 for a schematic description. ReUPred uses an iterative divide and conquer approach. Each iteration corresponds to a structural search, i.e., structural alignment of the query structure against all SRUL elements to identify a unit. The predicted unit corresponds to the aligned region in the query. At each cycle the algorithm forks (divides). 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. After the first cycle, i.e., after the "master" unit is found, SRUL is no longer used. Instead, a new ad hoc library is created on the fly. At the beginning of the second cycle, only the "master" unit populates the ad hoc library and all newly predicted units are included for search in the following cycles. The algorithm stops when the entire input protein is consumed, i.e., new input fragments are too short, or the structural search does not provide any new valid alignment. The predicted units are then collected and evaluated together (conquer). If the result does not satisfy a set of rules, the structural alignment filters for the "master" unit are relaxed and the entire iterative part is repeated from the beginning for up to four increasingly relaxed iterations. This strategy allows to predict both easy and difficult cases automatically. A valid solution for ReUPred is obtained when at least three units are found and their proximity in sequence is ensured by at

Table 1 Structural alignment constraints for the "master" unit

Iteration	TM-Score	RMSD (A)	Alignment (residues)	Unit gaps (%)
1	≥0.52	≤1.6	>21	<10
2	≥0.47	≤1.9	>17	<20
3	≥0.30	≤2.5	>16	<50
4	≥0.23	≤3.0	>14	<50

TM-Score and RMSD are the same provided by TM-Align. Coverage and gap are calculated as described in the manuscript. Different columns correspond to different algorithm runs that are performed on cascade until a valid solution is found

Table 3 Current RepeatsDB annotation of solenoid proteins

Class	Units	Detailed	Classified	Predicted
β	367	41	128	
α/β	180	19	70	
α	388	48	875	
Total	935	108	1073	7948

Units list the number of single defined repeat units. Detailed proteins have the unit position identified manually. Those protein for which the subclass assignment is known are classified, including "manually" and "by similarity". The predicted proteins are not yet classified

least one of two simple rules to measure unit proximity: (1) the total number of gaps between units is less than 40 residues, (2) the number of non-adjacent units divided by the total number of predicted units is less or equal to 0.25.

Replacing the original SRUL with an ad hoc library from the second cycle onward improves both computational cost and accuracy. SRUL is quite large, with 997 unit templates. Instead, the ad hoc library reaches the maximum size at the end of the algorithm and corresponds to the number of predicted units, drastically reducing the number of structural alignments. On the other hand, using only units from the query structure itself increases the accuracy as these are structurally more similar to each other than units from other proteins (data not shown). The class assignment is provided by simply reporting the classification assigned to the first "master" unit identified from SRUL.

ReUPred accuracy strongly depends on the quality of the structural alignments at each cycle. In particular, it is very important to correctly predict the first "master" unit because errors propagate. Alignments have to abide a set of rules and constraints that are much more stringent for the "master" search compared to successive cycles. Structural alignments are calculated using TM-Align (Zhang and Skolnick 2005), filtering by TM-Score, RMSD, alignment length and number of gaps. Tables 1 and 2 list all cutoff values for the cascaded four runs used to select valid alignments for the "master" and "secondary" units, executed on cascade until a valid solution is found. The parameters for structural alignments have been optimized manually on the training set to maximize the number of repeat proteins, for which a valid output is provided, and prediction accuracy, i.e., correct unit position assignment.

SRUL and datasets

The Structural Repeat Unit Library (SRUL) constitutes a fundamental part of the ReUPred input and represents the conformational space and diversity of bona fide repeat units. It has been generated by extracting all structural unit fragments from the "detailed" solenoid proteins in RepeatsDB (see Table 3 for statistics). After filtering units shorter than 10 residues and larger than 90, the solenoid SRUL is composed of 916 structural unit fragments from 108 different proteins non-redundant at the sequence level. After clustering the sequences with CD-HIT (Fu et al. 2012) at 40 % identity, 531 clusters are obtained. The largest cluster contains 17 units from 5 proteins and the others have less than 10 units each. From the structural point of view, SRUL is biased toward α-helical units. All-against-all structure similarity has been measured by TM-Align (Zhang and Skolnick 2005). Clustering at 0.6 TM-score generates 362 clusters, where the majority of α units (319) fall inside a single cluster.

Three different datasets have been used throughout this work. The training set has been generated from the

Table 2 Structural alignment constraints for the "secondary" units

Iteration	TM-Score	RMSD (A)	Alignment (residues)	Unit gaps (%)	Length ratio (%)
1	≥0.35	≤1.8	≤1.20	<40	≥70
2	≥0.30	≤2.0	≤1.15	<40	≥70
3	≥0.30	≤2.5	≤1.15	<40	≥70
4	≥0.30	≤3.0	≤1.10	<50	≥70

Columns are as in Table 2. The length ratio is calculated as the unit length divided by the length of the first "master" unit

"detailed" RepeatsDB entries (108 proteins) and represents the reference for unit prediction evaluation. Since SRUL has been generated from the same protein set, to benchmark ReUPred, all units coming from the target itself and all similar units (>30 % sequence identity) were removed from SRUL at each benchmarking step. Another set with all "classified" and "by similarity" entries (1075 proteins) has been used to test the ability to automatically classify repeat proteins and compare unit length prediction with RAPHAEL (Walsh et al. 2012). Finally, the dataset to test the detection of repeat proteins is taken from the same paper, i.e., 105 solenoid and 247 non-solenoid proteins with different topologies and no detectable sequence similarity.

Performance evaluation

For the unit-centric evaluation, a new strategy was implemented to take into consideration both the unit phase (position shift) and size. Predicted units are paired against the reference before defining the confusion matrix. Only one predicted unit is matched for each reference unit. When multiple predicted units overlap a single reference unit, the predicted unit with maximum overlap is selected. Figure 2 shows how a prediction is evaluated and how true positives (TP), false positives (FP), true negatives (TN) and false negatives (FN) are calculated. For classification, given a solenoid subclass, TP is the number of proteins with correct assignment, FN are proteins assigned the wrong class and FP are class assignments to wrong targets. TN is always zero since the test set contains only classified proteins. For all evaluations, the measures recall [or sensitivity; TP/(TP + FN)], precision [TP/(TP + FP)] and accuracy [(TP + TN)/(TP + FP + TN + FN)] are used. ReUPred is compared to the TAPO (Do Viet et al. 2015) and ConSole methods (Hrabe and Godzik 2014). TAPO predictions have been generated from the Web server (default parameters) considering only the first solution. ConSole predictions were generated locally by the stand-alone software (default parameters). The RAPHAEL period is provided in the RepeatsDB entry metadata. For all evaluations, ReUPred has been benchmarked after removing from SRUL units coming from the test protein or structurally similar units. The comparison with TAPO and ConSole was performed on a set of proteins for which all methods predict at least one unit, i.e., 89 out of 108 proteins.

Results

ReUPred was developed to predict both unit position and classify repeat proteins to automate the time-consuming manual annotation process of "detailed" annotation in RepeatsDB. See Fig. 3 for an example on plakophilin-1. Before benchmarking the main novel features, it is worthwhile to investigate whether ReUPred is able to correctly discriminate real repeats from non-repeat proteins. For this purpose, it has been compared with RAPHAEL (Walsh et al. 2012) on the original datasets (see Table 4). ReUPred correctly classifies 324 out of 352 domains (92 % accuracy). This is only somewhat lower than RAPHAEL on the same dataset (94.9 % and 95.7 %, for $S > 0$ and $S > 1$, respectively). A higher specificity could be obtained for ReUPred by setting a stronger filter on the last step of the algorithm, but that would affect coverage on the positive dataset. Even though ReUPred was designed to predict unit positions in tandem repeat proteins and not extensively optimized for repeat detection, this result demonstrates that the tool is also effective in discriminating repeat/non-repeat proteins.

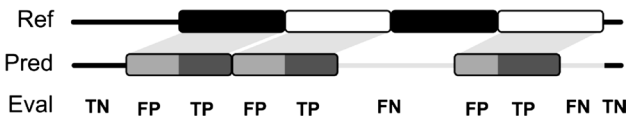

Fig. 2 Evaluation of repeat unit predictions. The evaluation (Eval) of a prediction (Pred) against a manually curated reference (Ref) is shown, with repeat units as *rounded rectangles*. The reference and predicted units are paired for maximal overlap. Residues are then categorized as true positives (TP), false positives (FP), true negatives (TN) and false negatives (FN) depending on repeat unit overlap

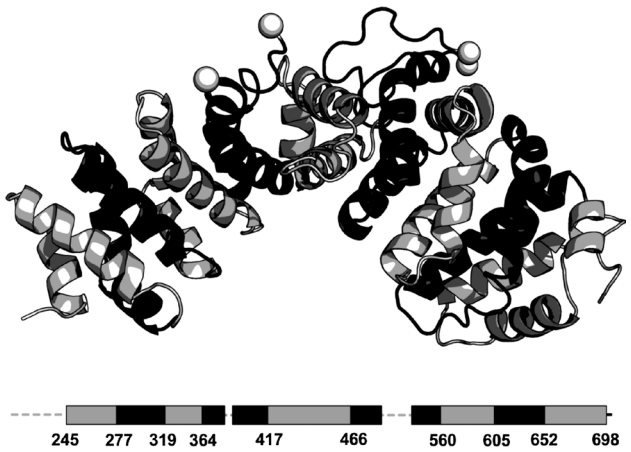

Fig. 3 ReUPred unit prediction for Plakophilin-1 (PDB code 1XM9, chain A). The structure is shown in cartoon representation in the top part with the schematized sequence below. Predicted units are represented in *black* and *gray*. *Dashed lines* represent missing residues in the PDB file (residues 388–396 and 481–508). The N- and C-terminal residues flanking the missing residues are shown as spheres in the structure

Repeat classification

ReUPred predicts units and fine classification for 83 % (893 proteins) of the RepeatsDB classified set. The class

Table 4 Solenoid detection performance on the RAPHAEL dataset

Method	TP	FP	TN	FN	Solenoids	Non-solenoids
RAPHAEL ($S > 0$)	94	7	240	11	89.5	97.2
RAPHAEL ($S > 1$)	91	1	246	14	86.7	99.6
ReUPred	81	4	243	24	77.1	95.3

The percentage of correctly classified solenoids and non-solenoids is shown together with the component true positives (TP), false positives (FP), true negatives (TN) and false negatives (FN). RAPHAEL is shown with the two SVM cutoff values as reported in the original paper

Table 5 ReUPred classification performance on the RepeatsDB classified dataset

Class	Recall	Precision	F-Measure	Accuracy
All-β	0.81	0.74	0.78	0.63
Mixed α/β	0.55	0.65	0.60	0.43
All-α	1.00	0.99	1.00	0.99
Total	0.94	0.94	0.94	0.89

See "Performance evaluation" for details on the measures used

assignment is obtained by simply transferring this information from the master unit found in SRUL. This approach has been proven to be effective as shown in Table 5. ReUPred works very well for the α class (III.3 in RepeatsDB). Instead, it is more difficult to correctly assign α/β and β examples. The low recall indicates that the cause of the problem is detecting units that do not have a good template in SRUL. This is an important result, as it indicates which RepeatsDB entries are worth manually annotating at the "detailed" level to improve ReUPred sensitivity and SRUL representation of the repetitive structural element universe. Low precision for β and α/β classes is due to a high number of false positive assignments. Looking at the data in detail, we found some ambiguous class assignments, e.g., PDB code 3ZYI, chain A, is annotated as α/β solenoid in RepeatsDB, but there are no helix elements except for a small fragment (residues 309–318) which is not repeated in the units. Since ReUPred predicts the class by transferring annotation from SRUL, if an SRUL element is misclassified the error propagates. ReUPred could be very useful to guide the manual refinement of RepeatsDB class annotations.

Unit prediction accuracy

ReUPred has been evaluated for unit prediction using the metric described in "Methods", i.e., penalizing predictions with a wrong phase or/and a wrong length. Table 6 shows a comparison with TAPO and ConSole in terms of predicted repeat residues on the detailed RepeatsDB set. The results are reported for each of the three main solenoid classes and for all proteins together. ReUPred always outperforms the other methods for all evaluation measures. In particular, the greatest improvement is observed for the α/β subclass, with an increase of 19 % accuracy compared with TAPO. The high accuracy for this class can be explained by the fact that mixed α/β units represent more structurally complex

Table 6 Comparative repeat unit prediction evaluation

Class	Method	Recall	Precision	*F*-Measure	Accuracy
All-β	TAPO	0.47	0.59	0.53	0.47
	ConSole	0.39	0.69	0.50	0.46
	ReUPred	**0.62**	**0.64**	**0.64**	**0.56**
Mixed α/β	TAPO	0.66	0.70	0.68	0.59
	ConSole	0.62	0.69	0.66	0.57
	ReUPred	**0.84**	**0.84**	**0.84**	**0.78**
All-α	TAPO	0.64	0.78	0.70	0.57
	ConSole	0.50	0.74	0.59	0.46
	ReUPred	**0.74**	**0.79**	**0.74**	**0.62**
Total	TAPO	0.58	0.70	0.64	0.53
	ConSole	0.48	0.71	0.58	0.49
	ReUPred	**0.71**	**0.75**	**0.73**	**0.62**

Performance evaluation is reported for each method on all RepeatsDB solenoid structures (All) and for the three subclasses separately (β, α/β and α). The best value for each quality measure is shown in bold. See "Performance evaluation" for details on the measures used

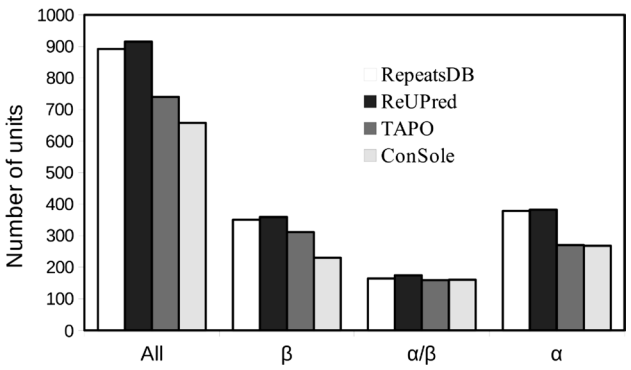

Fig. 4 The number of predicted units on the RepeatsDB detailed dataset. The manually curated reference (RepeatsDB) is shown next to the three prediction methods. ReUPred predicts more repeat units than the other two methods

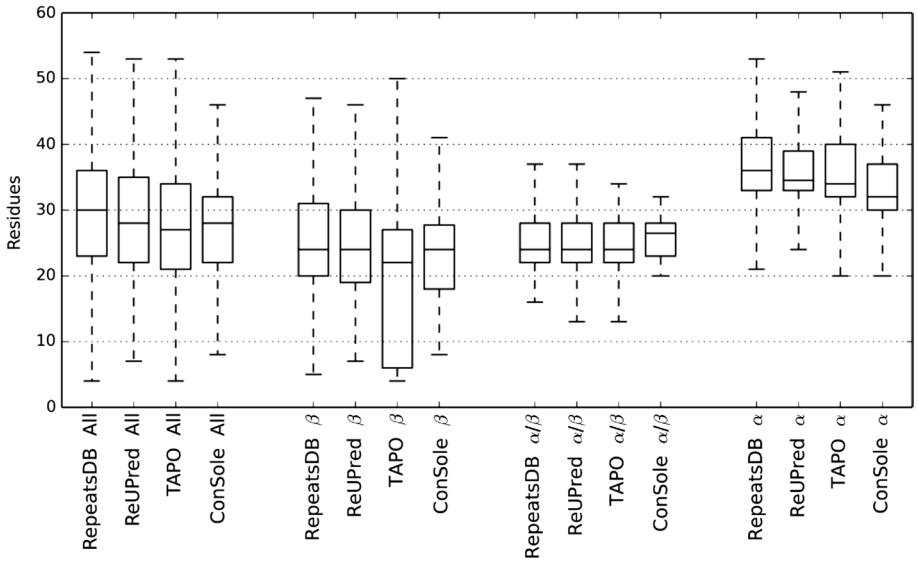

Fig. 5 Repeat unit periodicity *box plot* distribution on the RepeatsDB detailed dataset. The manually curated reference (RepeatsDB) is shown next to the three prediction methods

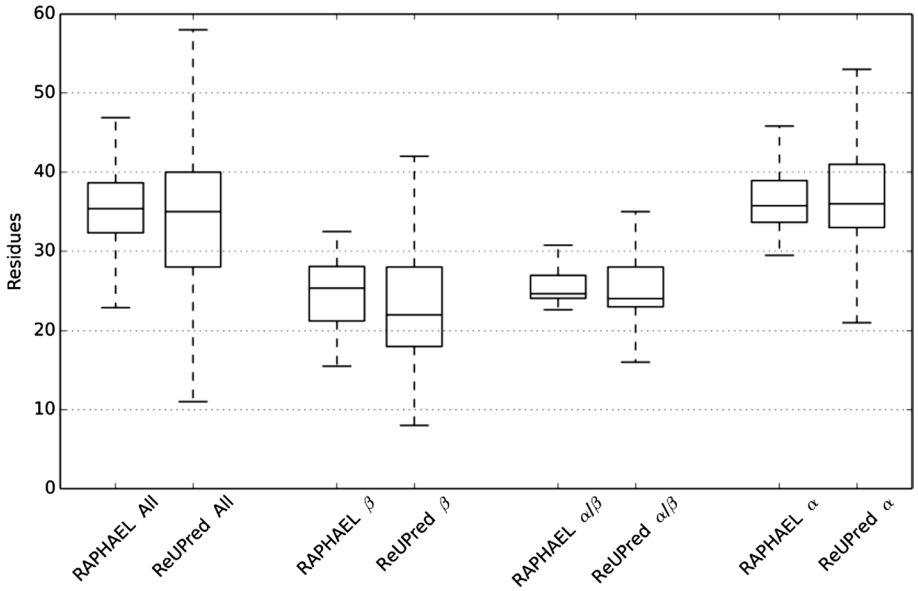

Fig. 6 Large-scale periodicity predictions on the RepeatsDB classified dataset. The original RAPHAEL periodicities are compared to ReUPred unit lengths as *box plot*

elements compared to all-α units. More information is coded in the structure unit, making it easier to discriminate wrong structural alignments. On the other hand, the most problematic subclass is all-β. Both recall and precision are lower for all methods compared with other subclasses. This may be explained by the fact that β solenoid units are more degenerated in the same protein than other solenoids and present a greater structural diversity with many insertions (data not shown). Moreover, they are shorter compared with all-α, generating worse structural alignments.

In addition to evaluating repeat annotations at the residue level, it is of interest to benchmark repeat units and their length distributions. Figure 4 shows the number of repeat units being identified by each method. Here again, ReUPred predicts more units than the other two methods. Both ConSole and TAPO generate units with the same size for a given structure and this may limit their ability to deal with insertions in solenoid proteins. ReUPred may therefore be better able to adapt to the irregular aspects of solenoid repeats. Figure 5 shows a box plot for the distribution of the predicted repeat periodicities against the RepeatsDB classified set. The median repeat length and standard deviations of ReUPred are very similar to the reference definition and on average match better than TAPO and ConSole. TAPO appears to underpredict the repeat length in β structures, probably because it also uses sequence information. ConSole on the other hand appears to have more difficulty with α-helices.

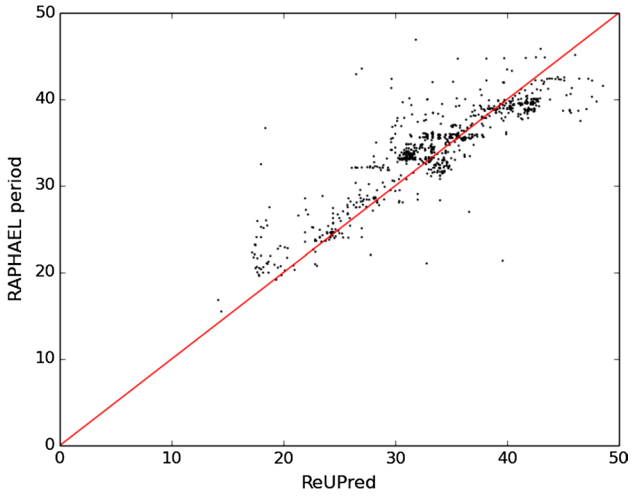

Fig. 7 Scatter plot of RAPHAEL and ReUPred periodicities on the RepeatsDB classified dataset. RAPHAEL produces a single periodicity per protein, whereas all predicted units were considered for ReUPred

Expanding the universe of known solenoids

Given the good performance of ReUPred for its intended purpose, i.e., classifying solenoid repeats and annotating their component units, it can be used to automatically expand the knowledge contained in RepeatsDB. The first step consists in establishing the baseline against the existing RAPHAEL annotations on the "classified" dataset. This contains annotations for solenoid class and predicted average repeat length. Since this dataset does not provide unit annotation, the simplest way to evaluate the performance is to compare the length of the predicted units with the repeat period predicted by RAPHAEL. This is the number of residues for which the symmetry signal is maximized, generating a single period for each protein. This is a big limitation, as it does not reflect the real situation where unit sizes vary inside a protein due to insertions which are frequent in solenoids. In particular, it is very relevant for the all-β class where almost all proteins have insertions. Figure 6 compares the distribution of ReUPred predicted unit length and RAPHAEL period for each solenoid class. Overall, both are very similar, with ReUPred having a wider range of periodicities as it is able to recognize irregularities in single repeat units. Only the distributions for all-β repeats differ more markedly. This class contains many structures with insertions which RAPHAEL struggles to summarize in a single fixed periodicity.

The scatter plot in Fig. 7 shows the correlation between the RAPHAEL period and ReUPred mean unit length calculated on each predicted protein. The two methods correlate strongly, with a Pearson correlation coefficient of 0.88 (P value $= 4.59 \times 10^{-290}$). On average, ReUPred predicts shorter units than the RAPHAEL period, 33.7 (SD 6.5) and 34.2 (SD 5.3) residues, respectively. When the RAPHAEL period is much larger (extreme points above the diagonal), ReUPred wrongly predicts two units instead of a single unit which would better represent the repetitive symmetry (e.g., PDB code 3L3F, chain X). For opposite cases, the contrary happens, i.e., ReUPred predicts a pair of units as a single element (e.g., PDB code 3PET, chain A).

To expand the annotation in RepeatsDB, ReUPred has been used to predict all repeat units for "classified" RepeatsDB solenoids. Since there is no comparison and no structural validation is possible, we chose to compare the annotation to Pfam. Figure 8 shows the very substantial increase in annotations both in terms of bona fide solenoid proteins and especially in the number of identified repeat units. The latter yields an increase of an order of magnitude compared to state-of-the-art sequence-based annotation in Pfam.

Conclusions

Classification and prediction of tandem repeat units are difficult problems currently addressed by expert manual curation.

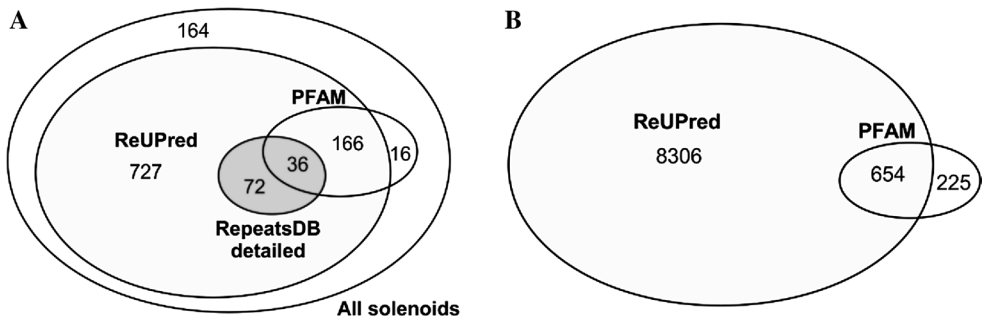

Fig. 8 Venn diagram of available annotations for RepeatsDB classified dataset. (**a**) Comparison of proteins with bona fide solenoid assignments. (**b**) The number of annotated repeat units in the dataset. The total number of repeat units in the dataset is unknown. ReUPred is able to increase the annotation by an order of magnitude in both cases

At the time of writing, it is available in RepeatsDB for only 3 % of the total putative repeat protein structures. ReUPred provides both the prediction of repetitive units and a finer classification in the RepeatsDB classification scheme. The algorithm works by exploiting a structure repeat unit library (SRUL) and an iterative decomposition of the input structure. While the performance was tested on the solenoid class, the method also works for other repeat types. ReUPred has been compared with other state-of-the-art methods, TAPO and ConSole, adopting an evaluation metric which takes into consideration both phase and size of the predicted units. Testing on a manually curated dataset obtained from the "detailed" RepeatsDB entries, ReUPred achieved the highest accuracy for all types of solenoids (β, α/β and α) with an overall increase of 9 % over TAPO and 13 % over ConSole. To provide an extended evaluation, a larger dataset with classified RepeatsDB entries without unit annotation was used. It was possible to test ReUPred ability of classifying solenoid structures and the correlation with periods predicted by RAPHAEL. ReUPred extended unit annotation and classification for almost all solenoids with high precision and accuracy. Moreover, the average unit length predicted by ReUPred strongly correlates with RAPHAEL, confirming the high quality of the predictions. Mixed α/β units are underrepresented in SRUL compared to the α and β classes, meaning that extending SRUL could yield a better recall and higher accuracy. ReUPred has also the ability to detect unit diversity inside a given target protein, recognizing fragment insertions that are not part of the repeat elements.

This work has demonstrated that repeat protein annotation can be made by repetitive template-based structural searches. Moreover, it shows that the approach can be applied reliably on a large scale, i.e., over all uncharacterized RepeatsDB entries, unveiling new scenarios for the analysis of the entire repeat protein universe.

Acknowledgments The authors are grateful to members of the BioComputing UP lab for insightful discussions. D.P. is funded by the FIRC project no. 16621.This project was partially supported by AIRC grant IG17753 and Elixir-Ita.

Compliance with ethical standards

Conflict of interest The authors declare no conflict of interest.

Research involving human participants and/or animals No.

References

Abraham A-L, Rocha EPC, Pothier J (2008) Swelfe: a detector of internal repeats in sequences and structures. Bioinformatics 24:1536–1537. doi:10.1093/bioinformatics/btn234

Andrade MA, Ponting CP, Gibson TJ, Bork P (2000) Homology-based method for identification of protein repeats using statistical significance estimates. J Mol Biol 298:521–537

Andrade MA, Petosa C, O'Donoghue SI et al (2001) Comparison of ARM and HEAT protein repeats. J Mol Biol 309:1–18. doi:10.1006/jmbi.2001.4624

Biegert A, Soding J (2008) De novo identification of highly diverged protein repeats by probabilistic consistency. Bioinformatics 24:807–814

Bazan JF, Kajava AV (2015) Designs on a curve. Nat Struct Mol Biol 22:103–105. doi:10.1038/nsmb.2966

Binz HK, Amstutz P, Kohl A et al (2004) High-affinity binders selected from designed ankyrin repeat protein libraries. Nat Biotechnol 22:575–582. doi:10.1038/nbt962

Björklund ÅK, Ekman D, Elofsson A (2006) Expansion of protein domain repeats. PLoS Comput Biol 2:0959–0970. doi:10.1371/journal.pcbi.0020114

Brunette TJ, Parmeggiani F, Huang P-S et al (2015) Exploring the repeat protein universe through computational protein design. Nature 528:580–584. doi:10.1038/nature16162

de Wit J, Hong W, Luo L, Ghosh A (2011) Role of leucine-rich repeat proteins in the development and function of neural circuits. Annu Rev Cell Dev Biol 27:697–729. doi:10.1146/annurev-cellbio-092910-154111

Di Domenico T, Potenza E, Walsh I et al (2014) RepeatsDB: a database of tandem repeat protein structures. Nucleic Acids Res 42:D352–D357. doi:10.1093/nar/gkt1175

Do Viet P, Roche DB, Kajava AV (2015) TAPO: a combined method for the identification of tandem repeats in protein structures. FEBS Lett 589:2611–2619. doi:10.1016/j.febslet.2015.08.025

Finn RD, Bateman A, Clements J et al (2014) Pfam: the protein families database. Nucleic Acids Res 42:D222–D230. doi:10.1093/nar/gkt1223

Fournier D, Palidwor GA, Shcherbinin S et al (2013) Functional and genomic analyses of alpha-solenoid proteins. PLoS One 8:e79894. doi:10.1371/journal.pone.0079894

Fu L, Niu B, Zhu Z et al (2012) CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinforma Oxf Engl 28:3150–3152. doi:10.1093/bioinformatics/bts565

Grove TZ, Cortajarena AL, Regan L (2008) Ligand binding by repeat proteins: natural and designed. Curr Opin Struct Biol 18:507–515. doi:10.1016/j.sbi.2008.05.008

Gruber M, Söding J, Lupas AN (2005) REPPER—repeats and their periodicities in fibrous proteins. Nucleic Acids Res 33:W239–W243

Heger A, Holm L (2000) Rapid automatic detection and alignment of repeats in protein sequences. Proteins 41:224–237. doi:10.1002/1097-0134(20001101)41:2<224:aid-prot70>3.0.co;2-z

Höcker B (2014) Design of proteins from smaller fragments—learning from evolution. Curr Opin Struct Biol 27:56–62. doi:10.1016/j.sbi.2014.04.007

Hrabe T, Godzik A (2014) ConSole: using modularity of Contact maps to locate Solenoid domains in protein structures. BMC Bioinformatics 15:119. doi:10.1186/1471-2105-15-119

Jorda J, Kajava AV (2009) T-REKS: identification of Tandem REpeats in sequences with a *K*-meanS based algorithm. Bioinformatics 25:2632–2638

Kajava AV (2001) Review: proteins with repeated sequence–structural prediction and modeling. J Struct Biol 134:132–144. doi:10.1006/jsbi.2000.4328

Kajava AV (2012) Tandem repeats in proteins: from sequence to structure. J Struct Biol 179:279–288. doi:10.1016/j.jsb.2011.08.009

Kim M, Abdi K, Lee G et al (2010) Fast and forceful refolding of stretched α-helical solenoid proteins. Biophys J 98:3086–3092. doi:10.1016/j.bpj.2010.02.054

Kobe B, Kajava AV (2000) When protein folding is simplified to protein coiling: the continuum of solenoid protein structures. Trends Biochem Sci 25:509–515

Marcotte EM, Pellegrini M, Ng H-L et al (1999a) Detecting Protein Function and Protein-Protein Interactions from Genome Sequences. Science 285:751–753. doi:10.1126/science.285.5428.751

Marcotte EM, Pellegrini M, Yeates TO, Eisenberg D (1999b) A census of protein repeats. J Mol Biol 293:151–160. doi:10.1006/jmbi.1999.3136

Mistry J, Coggill P, Eberhardt RY et al (2013) The challenge of increasing Pfam coverage of the human proteome. Database 2013. doi:10.1093/database/bat023

Mitchell A, Chang H-Y, Daugherty L et al (2015) The InterPro protein families database: the classification resource after 15 years. Nucleic Acids Res 43:D213–D221. doi:10.1093/nar/gku1243

Newman AM, Cooper JB (2007) XSTREAM: a practical algorithm for identification and architecture modeling of tandem repeats in protein sequences. BMC Bioinforma 8:382

Paladin L, Tosatto SCE (2015) Comparison of protein repeat classifications based on structure and sequence families. Biochem Soc Trans 43:832–837. doi:10.1042/BST20150079

Park K, Shen BW, Parmeggiani F et al (2015) Control of repeat-protein curvature by computational protein design. Nat Struct Mol Biol 22:167–174

Parmeggiani F, Pellarin R, Larsen AP et al (2008) Designed armadillo repeat proteins as general peptide-binding scaffolds: consensus design and computational optimization of the hydrophobic core. J Mol Biol 376:1282–1304. doi:10.1016/j.jmb.2007.12.014

Pellegrini M (2015) Tandem repeats in proteins: prediction algorithms and biological role. Front Bioeng Biotechnol. doi:10.3389/fbioe.2015.00143

Pellegrini M, Renda ME, Vecchio A (2012) Ab initio detection of fuzzy amino acid tandem repeats in protein sequences. BMC Bioinformatics 13:1–13. doi:10.1186/1471-2105-13-S3-S8

Sabarinathan R, Basu R, Sekar K (2010) ProSTRIP: a method to find similar structural repeats in three-dimensional protein structures. Comput Biol Chem 34:126–130. doi:10.1016/j.compbiolchem.2010.03.006

Schaper E, Korsunsky A, Messina A et al (2015) TRAL: Tandem repeat annotation library. Bioinformatics btv306. doi:10.1093/bioinformatics/btv306

Söding J, Remmert M, Biegert A (2006) HHrep: de novo protein repeat detection and the origin of TIM barrels. Nucleic Acids Res 34:W137–W142. doi:10.1093/nar/gkl130

Szklarczyk R, Heringa J (2004) Tracking repeats using significance and transitivity. Bioinformatics 20:i311–i317

Varadamsetty G, Tremmel D, Hansen S et al (2012) Designed Armadillo repeat proteins: library generation, characterization and selection of peptide binders with high specificity. J Mol Biol 424:68–87. doi:10.1016/j.jmb.2012.08.029

Walsh I, Sirocco FG, Minervini G et al (2012) RAPHAEL: recognition, periodicity and insertion assignment of solenoid protein structures. Bioinformatics 28:3257–3264. doi:10.1093/bioinformatics/bts550

Zhang Y, Skolnick J (2005) TM-align: a protein structure alignment algorithm based on the TM-score. Nucleic Acids Res 33:2302