Bioinformatics, 32(13), 2016, 1952–1958
doi: 10.1093/bioinformatics/btw118
Advance Access Publication Date: 7 March 2016
Original Paper

OXFORD

Sequence analysis

Tally: a scoring tool for boundary determination between repetitive and non-repetitive protein sequences

François D. Richard[1,2], Ronnie Alves[2,3] and Andrey V. Kajava[1,2,4,*]

[1]Centre de Recherche en Biologie cellulaire de Montpellier (CRBM), UMR 5237 CNRS, Université Montpellier 1919 Route de Mende, Cedex 5, Montpellier 34293, France, [2]Institut de Biologie Computationnelle (IBC), Montpellier 34095, France, [3]Pós-Graduação em Ciência da Computação (PPGCC), Universidade Federal do Pará, Belém, Brazil and [4]University ITMO, Institute of Bioengineering, St. Petersburg 197101, Russia

*To whom correspondence should be addressed.
Associate Editor: Burkhard Rost

Received on October 26, 2015; revised on February 11, 2016; accepted on February 25, 2016

Abstract

Motivation: Tandem Repeats (TRs) are abundant in proteins, having a variety of fundamental functions. In many cases, evolution has blurred their repetitive patterns. This leads to the problem of distinguishing between sequences that contain highly imperfect TRs, and the sequences without TRs. The 3D structure of proteins can be used as a benchmarking criterion for TR detection in sequences, because the vast majority of proteins having TRs in sequences are built of repetitive 3D structural blocks. According to our benchmark, none of the existing scoring methods are able to clearly distinguish, based on the sequence analysis, between structures with and without 3D TRs.
Results: We developed a scoring tool called *Tally*, which is based on a machine learning approach. *Tally* is able to achieve a better separation between sequences with structural TRs and sequences of aperiodic structures, than existing scoring procedures. It performs at a level of 81% sensitivity, while achieving a high specificity of 74% and an Area Under the Receiver Operating Characteristic Curve of 86%. *Tally* can be used to select a set of structurally and functionally meaningful TRs from all TRs detected in proteomes. The generated dataset is available for benchmarking purposes.
Availability and implementation: Source code is available upon request. Tool and dataset can be accessed through our website: http://bioinfo.montp.cnrs.fr/?r=Tally.
Contact: andrey.kajava@crbm.cnrs.fr
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Today, the growth of the protein sequence data significantly exceeds the capacity to analyze this data. Many proteins contain periodic sequences representing arrays of repeats that are directly adjacent to each other (Andrade *et al.*, 2001; Heringa, 1998; Kajava, 2012), so called tandem repeats (TRs). TRs occur in at least 14% of all proteins (Marcotte *et al.*, 1999). Moreover, they are found in every third human protein. Highly divergent, they range from a single amino acid repetition to repeated domains of more than 100 residues. Numerous studies demonstrated the fundamental functional importance of such TRs and their involvement in human diseases (Liggett and Sidransky, 1998; Morin, 1999; Paladin *et al.*, 2015; Simeonova *et al.*, 2012). Thus, TR regions are abundant in proteomes and are related to major health threats in modern society. Along this line, understanding of their sequence–structure–function relationship and evolutionary mechanisms is a promising path in the identification of targets for new medicines and vaccines.

Over the course of evolution, perfect TRs tend to degenerate, as a result of mutations, substitutions and deletions. Therefore, the main problem is to distinguish between sequences that contain

highly imperfect TRs and aperiodic sequences without TRs (Fig. 1). To solve this problem a number of computational tools have been developed, which detects TRs in protein sequences (e.g. Biegert and Söding, 2008; Jorda and Kajava, 2009; Newman and Cooper, 2007; Szklarczyk and Heringa, 2004). A survey of these methods shows that no best approach to score putative TRs in sequences, exists that covers the whole range of repeats (Kajava, 2012). This evokes the necessity to integrate several algorithms into a pipeline (Richard and Kajava, 2014). Within this pipeline, it is also instrumental to have a universal score, to determine the boundary between repetitive and non-repetitive protein sequences, as at present each of these computational programs use their own measure.

For example, most algorithms consider multiple sequence alignments (MSA) of repeats as a collection of columns, such scores compute an intermediate score over each column and obtain the global score by averaging the intermediates. *Psim* (Jorda and Kajava, 2009) considers the frequency of the most common element in a column, *Sdiff* (Schaper *et al.*, 2012) counts the number of mutations observed in a column, Shannon entropy (*S*) (Valdar, 2002) accounts for the frequencies of each amino acids in a column. Other metrics include evolutionary information such as *P-value-phylo* (Schaper *et al.*, 2012). Several scores are embedded into their original TR predictors such as TRUST (Szklarczyk and Heringa, 2004), *pftools* (Bucher *et al.*, 1996) or HHrepID (Biegert and Söding, 2008). None of these methods, apart from HHrepID, use information about 3D protein structure, to set the boundary between repetitive and non-repetitive protein sequences.

The TR scoring functions can be tested on the structural definition of TRs. For the TR sequences that fold into a stable 3D structure, one can define the criteria for 'true' TRs, as an existence of repetitive blocks in structure. Indeed, the 3D structure of the folded TRs, does not allow mutations that destabilize the structure. It is important to mention that such TRs correspond to repetitive structural blocks (Kajava, 2012). Sometimes, but less frequently, functional constraints also play a role in the maintenance of repetitive sequence patterns during evolution. Thus, TR scoring functions can be tested against a dataset of the 'true' TRs found both in sequence and in structure (TR-SS) and 'false' TRs only found in sequence but not in the structure (TR-SNS).

Recently, given this structural definition of TRs, we demonstrated (Richard and Kajava, 2015) that none of the existing scoring metrics described above, were able to accurately distinguish between TR-SS and TR-SNS. Here, we present a new method called *Tally*, for scoring TRs, based on a robust Machine Learning (ML) approach trained on an annotated TR-SS and TR-SNS datasets. To the best of our knowledge this is the first attempt to benchmark scoring systems that aim to find a proper boundary between TR-SS and TR-SNS. ML features includes some of the previously used scoring metrics and several other characteristics, generating an optimal scoring combination. Our approach performs significantly better than existing approaches, over the complete range of TRs.

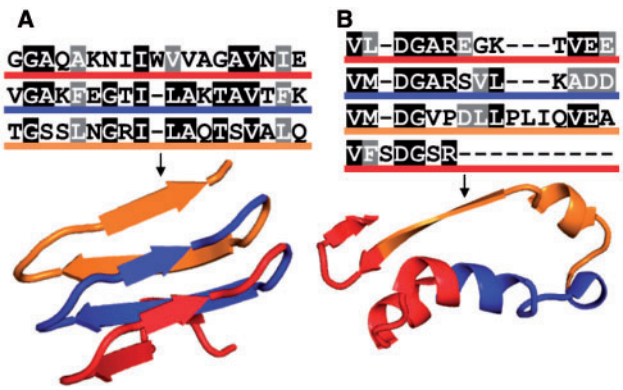

Fig. 1. Examples of proteins where TRs found in sequence either correspond to (**A**) presence (PDB code 3VN3 (Kondo *et al.*, 2011)) or (**B**) absence (4FUR) of TRs in the structure. TRs found in the sequence are mapped to their structures. The sequence repeats are colored by a repetitive pattern

2 Methods

The development of the scoring tool was divided into 3 steps (Fig. 2):

(1) Construction of datasets of TR-SS and TR-SNS
(2) Machine Learning (ML) feature engineering
(3) ML model optimization

2.1 Building datasets of TR-SS and TR-SNS

We build the datasets by taking into account proteins from two sources: (i) the 321 manually annotated proteins with TRs in structure (3D TRs) from RepeatsDB (Di Domenico *et al.*, 2014) and (ii) 340 proteins not covered by RepeatsDB. The proteins were selected as follows: first, we clustered the proteins from the PDB at 100% sequence identity using CD-HIT (Li and Godzik, 2006). Second, we identified TRs in the sequences of all these proteins by using HHrepID (Biegert and Söding, 2008), TRUST (Szklarczyk and Heringa, 2004) and T-REKS (Jorda and Kajava, 2009) using their default parameters. HHrepID has been run on the output from HHblits (Remmert *et al.*, 2012), with default parameters, and three iterations on the non-redundant database from NCBI clustered at 20% identity. Among these TRs in sequences, we selected only those that correspond to the structural regions without residues that we annotated in the PDB as 'MISSING' (remark 465). Among the proteins that do not belong to RepeatsDB we selected only TR

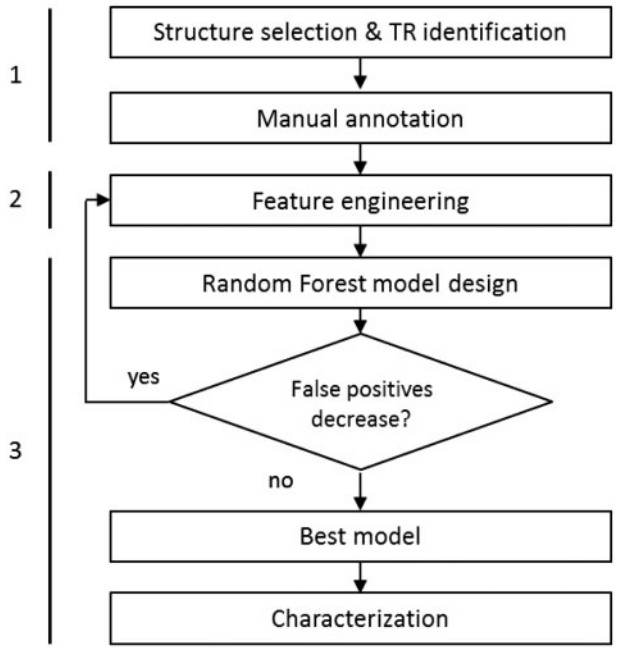

Fig. 2. Flowchart used to build the *Tally* scoring tool

sequences with scores of *P-value-phylo* <0.001 (Schaper *et al.*, 2012) to obtain a more confident dataset. The TR sequences were then clustered at 40% sequence identity by CD-HIT (Li and Godzik, 2006) in order to decrease structural redundancy, while preserving structural diversity. The threshold of 40% was chosen because proteins having this sequence identity, with a high probability have similar structures (Chothia and Lesk, 1986).

At the next step, the TRs from RepeatsDB and the newly selected TRs were combined. Then, in order to obtain a non-redundant dataset of the sequence TRs we removed overlapping TRs, accordingly to the following algorithm: (i) if 2 TRs overlap, we consider the number of residue positions corresponding to gaps in their MSAs and keep the MSA having the lease number of gap positions; (ii) if the numbers of gapped positions are the same, we select the TRs having more repeat units; (iii) if the number of repeat units is equal, we take the TR with the longest repeat unit. As a result, we found 241 TRs in the sequences of the 321 proteins, from RepeatsDB, and 341 TRs in the sequences of the 340 additional proteins from the PDB.

We verified that none of the TRs were identical, defining a TR by the number of repeat units, its repeat unit length, and its sequence. We also evaluated redundancy of the TR sequences using CLUSTALW2 (Larkin *et al.*, 2007), with default parameters, obtaining an average score for pairwise alignment of 8.5% identity.

Each selected TR has been mapped on to its structure and classified either as TR-SSs or TR-SNSs, with the aid of TAPO (Do Viet *et al.*, 2015), a program for identifying structural repeats. Subsequently, the TAPO results were also manually verified. The criteria used in TAPO to define the 3D structure as repetitive were the following: the repeats detected by the sequence analysis correspond to: (i) the structural segments with similar distribution of secondary structure (based on DSSP data (Touw *et al.*, 2015)); (ii) that have similar super-secondary arrangements (based on TM-score (Zhang, 2005)); and (iii) similar inter-repeat interactions (based on map of contacts between residues). As a result, we obtained 441 TR-SSs and 141 TR-SNSs. For each TR, we recorded the average length of the repeat unit and its secondary structure type. We also annotated the TRs in respect of whether or not they are in the core of the structure. Based on the visual inspection, TR was considered to be in the structural core if this region cannot be removed from the structure without destroying its structural integrity. Non-core TRs are frequently located in long loops.

The dataset is available upon request or at http://bioinfo.montp.cnrs.fr/?r=Tally. It can be used by other researchers as a benchmark to set boundary between TR-SS and TR-SNS.

2.2 Machine Learning (ML) feature engineering

We generated 5 ML feature families containing in total 32 features (Supplementary Table S1). They are:

(1) Sequence family (4 features): the known scores (*Psim*, *Sdiff*, *S* and *P-value-phylo*) that can be directly applied to a MSA of repeats (reviewed in Richard and Kajava, 2015).
(2) Substitution matrix family (18 features): includes scores of the Sequence family that take into account substitution matrices BLOSUM62, BLOSUM45 and PAM120. It also includes a variant where an extra penalty for gaps has been included in the matrix.
(3) Physiochemical properties family (3 features): gathers the scores of the first family applied on a modified MSA, where residues are replaced by their physicochemical properties: L, V, I, M as hydrophobic, F, Y, W as aromatics, D, E, K, R as charged, Q, H, N, S, T as hydrophilic and A, C, G, P were considered independently.
(4) Gap family (4 features): includes features related to the gaps in the MSA, such as the standard deviation of the repeat unit length, number of gaps in the alignment as well as number of residues (both normalized by the total number of characters in the MSA). It also includes a variant of the *Sdiff* score on protein sequence (the best among the other individual scores (Fig. 3)) that includes an extra penalty for gaps.
(5) Secondary structure family (3 features): includes the scores of the sequence family, with residues of the MSA replaced by their predicted secondary structure. Recent secondary structure prediction tools rely on ML approaches, where the learning set could include some of the structures also used in our datasets. Therefore, to avoid any bias we used a ML free method for secondary structure prediction. This method is described in (Williams *et al.*, 1987) and based on an optimized version of the Chou and Fasman algorithm (Chou and Fasman, 1974).

We evaluated the individual contribution of each family, which offer similar generalization power, discriminating TR-SSs and TR-SNSs at the level of about 70% of AUC (Supplementary Fig. S1A). Feature importance was evaluated (i) locally, by means of the learning function (the mean of Gini index reduction was employed by Random Forest (RF)) and (ii) globally, by means of AUC scores and the respective RF's models.

2.3 ML model optimization and characterization

We used a typical ML methodology to build our classification model (Flach, 2012; Walsh *et al.*, 2015). We defined a basic

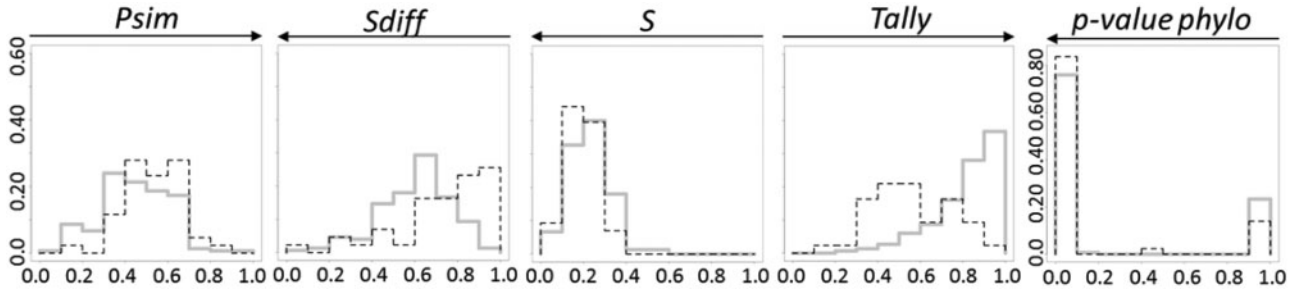

Fig. 3. Power of distinction between TR-SSs and TR-SNSs by currently used scores and *Tally*. Histograms of the score distributions are normalized by the total number of the TRs in each category (441 TR-SSs and 141 TR-SNSs). The grey line represents the score distributions of TR-SSs, the dashed black line represents the score distributions of TR-SNSs. Entropy (*S*) is normalized between 0 and 1. The y-axis of first 4 histograms share the same scale. Arrows above the histograms indicate the direction of the increase of TR perfection. A good distinction between TR-SSs and TR-SNSs would be when these distributions do not overlap and the TR-SSs are shifted toward the perfect TRs, and the TR-SNSs are at the opposite end

protocol as follows: (i) to select a dataset and features, (ii) to randomly split the dataset, (iii) to undertake a learning phase with 2/3 of the dataset, grid search to optimize parameters, 10-fold cross validation and model selection based on AUC and (iv) to run the testing phase on the remaining data, computing AUC, specificity and sensitivity.

The feature space cannot be explored efficiently by linear classifiers, thus the ensemble learning strategies, such as RF, can be more effective in finding a defined boundary between TR-SSs and TR-SNSs. Moreover, experimental evaluation with other ML strategies highlighted the robustness of the Random Forest (RF) approach when applied to biological sequences (Chen and Ishwaran, 2012; Mendoza *et al.*, 2013). Therefore, we choose RF as the primary approach to test. The testing gave promising results on our data (Supplementary Fig. S1A), providing us with a reason for further RF model optimization. The grid search optimized the number of trees (10, 100, 500, 1000), the maximal depth of the trees (5, 10, 50) and the maximal features used to characterize a node in the tree (from 1 to the number of feature used in the protocol).

The initial split of the dataset to training and testing sets, was carried out to guarantee an equal proportion of TR-SSs and TR-SNSs in both training and testing sets.

Two feature selection approaches were used: (i) a 'family independent' approach where the features are selected exclusively according to their importance and (ii) a 'family dependent' approach, where each feature family must be represented at least once in the feature selection. The best results have been obtained using the 'family dependent' approach. The process of optimization can be described as follows. First, we ran the basic protocol, with the entire dataset and all the features (Supplementary Fig. S1C, D). Second, we selected the best feature in each family or a top *k* model, where *k* is a number of top features either covering the 5 families, or covering the first pool of features according to their importance. The basic protocol was rerun 10 times with the complete dataset, only using the best features. Third, we selected the best models and rerun 10 times the basic protocol, omitting the split, in order to repeat the cross-validation. Once we have determined the best model we identify cases that are predicted as TR-SS but assigned as TR-SNS. This led to a double check of class assignment in the dataset, aiming to evaluate whether the count of false positives can be reduced by adding or removing features. With the new set of features, we reran the complete process, designing an iterative model building phase. The risk of over fitting has been monitored by analyzing the difference of performance the between training and testing phases.

The optimization process revealed a single best feature for each of the following families: 'raw sequences' (*Sdiff*), 'physicochemical properties' (*Sdiffeq*) and 'gaps' (*msastd*). For the secondary structure family two features were ranked with the same importance (*SdiffSS* and *PsimSS*). Feature selection for the substitution matrix family was less obvious (Supplementary Fig. S1C). Therefore, as a final optimization step we decided to use the five best pre-selected features mentioned above and to test all the features in the substitution matrix family, one by one, to find the best top six feature model. Our protocol spotted '*Sdiffb62*' as the best feature. Finally, the feature '*SdiffGap*' appear in the top of feature importance (Supplementary Fig. S1C) and its addition allowed us to obtain the best performance with a seven features model (Fig. 4).

The machine learning approach has been undertaken using the scikit-learn 0.16.0 package (Pedregosa *et al.*, 2011) available on Python 3.3.2. Source code of *Tally* partially rely on scripts used in (Schaper *et al.*, 2012).

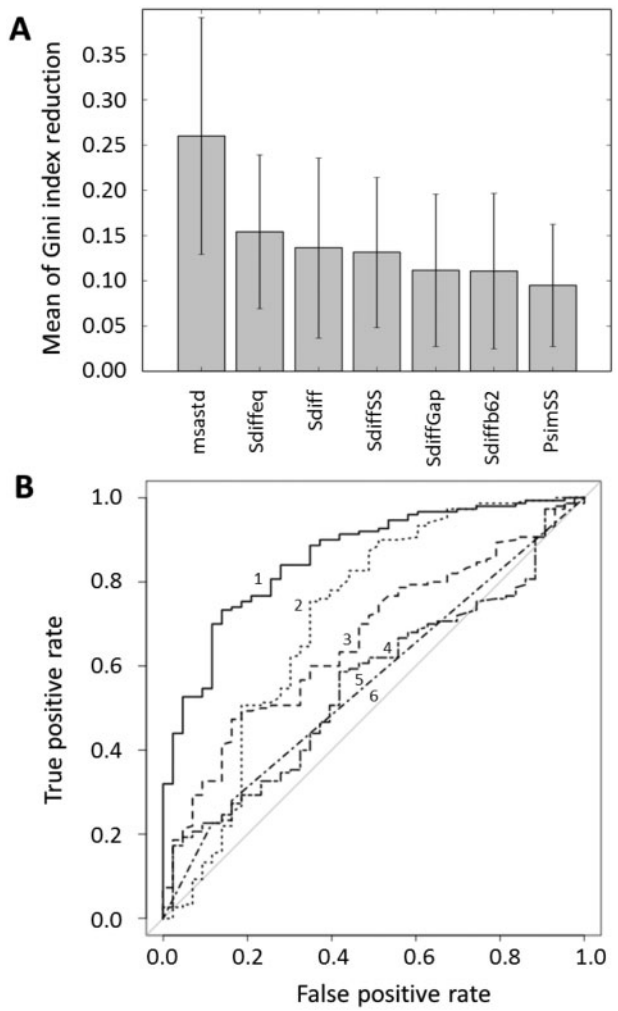

Fig. 4. Characteristics and performance of the machine learning approach after considering the best 7 features (*Sdiff, Sdiffeq, msastd, SdiffSS, PsimSS, Sdiffb62, SdiffGap*) that make up *Tally*. **A**: Feature importance obtained with *Tally*. The feature importance, or Gini index importance, is based on the node impurity measure for node splitting. The importance of a variable is defined as the Gini index reduction for the variable summed over all nodes for each tree in the forest, normalized by the number of trees. **B**: ROC curves obtained on the testing set used by *Tally* (1) and the other scoring methods: random (6), *S* (4), *P-value-phylo* (5), *Psim* (3) and *Sdiff* (2). Their AUCs are 0.86, 0.5, 0.56, 0.56, 0.66, 0.71, respectively

2.4 Other classifiers

Other classifiers have been tested with 10-fold cross validation and 10 repetitions. The grid search for KNN optimized the number of neighbors (2, 5, 10, 50) and the algorithm to find the optimum parameter ('auto', 'ball_tree', 'kd_tree', 'brute'). The grid search for SVM (with RBF kernel) optimized C and gamma parameters with logarithm scale from 10^{-2} to 10^{8} and from 10^{-5} to 10^{3} respectively. C parameter controls the cost of misclassification on the training data whereas gamma is a parameter of the kernel. Small values for those parameters lead to a high tolerance for misclassification of cases, while high values tend to reduce those cases but can lead to over fitting of the data. The grid search allows us to find the best compromise. A GLM classifier using the logistic regression algorithm has also been used. Comparative performance is given in Supplementary Figure S1B showing that RF was the best approach in both performance and robustness.

3 Results and discussion

Tally is based on the best model obtained during the model optimization phase. It uses 7 features (*Sdiff*, *Sdiffeq*, *msastd*, *SdiffSS*, *PsimSS*, *Sdiffb62*, *SdiffGap*) and gives an Area Under a ROC Curve (AUC) of 0.86 (best at 1), a specificity of 0.74, and a sensitivity of 0.81 for a threshold of 0.71 on the testing set (Fig. 4B). Here, we defined sensitivity as TP/(TP + FN) and specificity as TN/(TN + FP), where TP (True Positive) are the corrected predicted TR-SSs; TN (True Negative) the correctly predicted TR-SNSs; FP (False Positive) the TR-SNSs predicted as being TR-SSs; and FN (False Negative) are the TR-SSs predicted as TR-SNSs. The thresholds for the benchmarked methods that were used for evaluation of their sensitivities and specificities, were established as the point given the best balance of both parameters on the ROC curves (Fig. 4B). As recommended in (Walsh *et al.*, 2015) we checked the generalization power of *Tally* by comparing its performance on training and testing data. The averaged AUCs on training and testing data obtained by the 10 repetitions of the 10-fold cross validations are 0.77 and 0.85 respectively (with a standard deviation of 0.013 and 0.012 respectively and a *P*-value for Shapiro-Wilk normality test of 0.0015, and 0.3949 respectively). The fact that the AUC on the training set is significantly smaller (Wilcoxon signed-rank test, *P*-value <0.0001), than the AUC on the testing data confirms a good generalization power of the model while the low variability of the AUC over the repetitions attests the robustness of the model.

3.1 Tally compared to other scoring methods

3.1.1 Tally and the non-embedded scoring methods

The performance of *Tally* represents a significant improvement of the power of discrimination, compare to the other scoring method that use sequence information only (Figs 3, 4B). Indeed, the other methods gave the following values of AUC, specificity, sensitivity *Psim* (0.66, 0.25, 0.49), *Sdiff* (0.71, 0.51, 0.83), *S* (0.56, 0.33, 0.63), *P-value-phylo* (0.56, 0.16, 0.75) respectively.

3.1.2 Tally and the embedded scoring methods

We also tested the performance of *Tally* against TRUST and HHrepID, which includes protein homology information in their scoring procedures. Since TRUST and HHrepID embed their scoring function and are able to score only the MSA they produce, we have to compare the performance of *Tally* on the output of each prediction algorithm separately. We built a new dataset, which includes 337 and 506 TRs found by HHrepID and TRUST respectively, from the proteins of our reference dataset. Analyzing the 3D structure of these TRs, we assigned each protein to either TR-SS or TR-SNS. It is worth mentioning that some of these TRs were unseen by *Tally* during the ML.

Among the 337 TRs determined by HHrepID, 311 are TR-SSs and 26 are TR-SNSs. Applied on 337 TRs from the HHrepID output, *Tally* correctly assigned 236 TR-SSs and incorrectly predicted 16 TR-SNS as TR-SSs (Fig. 5A).

Among the 506 TRs determined by TRUST, 350 are TR-SSs and 156 TRs are TR-SNSs. Applied on 506 TRs from the TRUST output, *Tally* correctly predict 241 TR-SSs and incorrectly predict 38 TRs, which are TR-SSs while, in reality, they are TR-SNSs (Fig. 5B). Therefore, applied on HHrepID and TRUST output separately, *Tally* unveils both less TPs and FPs. Nevertheless, the main strength of *Tally* is its possibility to be applied on any MSA, along with its ability to be integrated into a Meta Repeat Finder pipeline, as described in (Richard and Kajava, 2014). In particular, it is possible to use *Tally* simultaneously on both HHrepID and TRUST outputs.

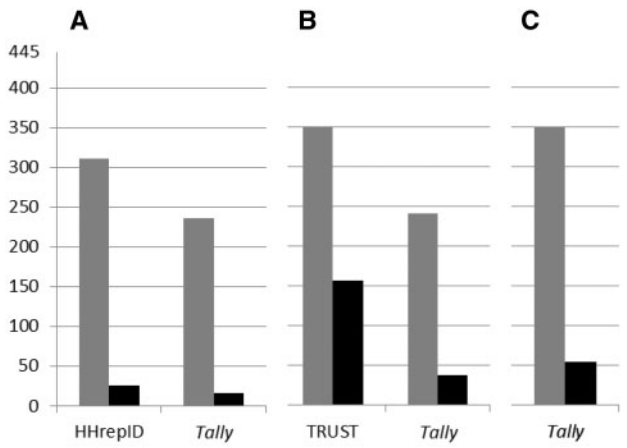

Fig. 5. Comparison of *Tally* and the embedded scoring methods. Grey bars in **A** and **B** correspond to the number of TR-SSs detect by each predictor (True Positive-TP) and in C it is the number of non-overlapping 3D TR areas detect by *Tally* from the 445 TRs of the combined HHrepID and TRUST dataset. In black is the number of TR-SNSs predicted as TR-SSs (False Positive – FP). (**A**) Comparison of HHrepID and *Tally* performance on the output of HHrepID. (**B**) Comparison of TRUST and *Tally* performance on the output of TRUST. (**C**) Performance of *Tally* on the combined output of HHrepID and TRUST

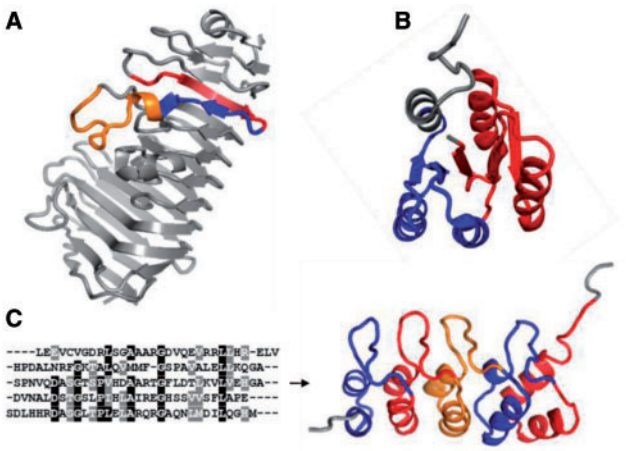

Fig. 6. Examples of misclassified cases. The TRs found in sequence are mapped to the structure. Repeat units are colored in structure by a repetitive pattern. The grey parts of the structure have not been identified as involved in TRs in sequence. (**A**) TR-SNS not in the core of the protein (1K5C (Shimizu *et al.*, 2001)). (**B**) TR-SNS where repeat units does not share the same interactions (2L69). (**C**) TR-SS (1AP7 (Luh *et al.*, 1997)) not validated by *Tally* because it is a highly degenerated TRs

Among the total of 843 (337 + 506) TRs found by HHrepID and TRUST we detected 445 non-overlapping TR-SSs and 164 non overlapping TR-SNSs. Ideally, the scoring procedure should identify all of the 445 structural repeats and none of the 164 TR-SNSs. *Tally* unveils 349 3D TRs and incorrectly predicts as TR-SSs 50 TR-SNSs (compared with HHrepID – 311 and 26, TRUST – 344 and 156) (Fig. 5). The Matthews Correlation Coefficient (MCC) that is frequently used to evaluate the quality of predictions (Matthews, 1975) is equal to 0.44 for *Tally*, 0.48 for HHrepID and −0.21 for TRUST. Taking into account that the closer MCC to 1 the better is the prediction quality, we can conclude that *Tally* performance is comparable with HHrepID. At the same time *Tally*'s output is the most complete with the sensitivity 0.78 for *Tally* and 0.70 for HHrepID.

Fig. 7. Example of a misclassified case that can be rescued by MSA optimization. The original MSA is given by TRUST on the chain B of the PDB code 4EIV (Tonkin *et al.*, 2012). *Tally* gives a score of 0.64 which is not large enough to be considered as TR-SS. After removing the third repeat unit and aligning the two remaining ones with MUSCLE (Edgar, 2004) the *Tally* score of this MSA becomes 0.75 which is above the threshold to be considered as a TR-SS and closer to the actual structural alignment

3.2 Exploring *Tally*'s limitations

Tally has been applied to the whole dataset, including training and testing data. As a result, a large amount of negative and positive candidates (119 among 141, and 368 among 441) are well classified. In the next section we take a closer look at *Tally* misclassification to understand its limitations.

3.2.1 TR-SNS predicted as TR-SS

Among the 141 TR-SNS from our benchmark dataset only 22 cases are misclassified by *Tally*. Analysing these proteins we found that in 9 of the 22 cases, TRs in sequence correspond to the peripheral (non-core) part of the 3D structures (Fig. 6A). This suggests an evolutionary scenario in which such TR region bulged out from the loop region. The initial conformation of the repeats within the loop may be labile by itself while the fixed conformation can be induced by the core structure. In this context, since the emergence, each of the repeats may adopt different conformations. The observed non-TR structures predetermine disappearance of these TRs in the sequence during evolution. The other 5 of the 22 proteins do not have 3D TRs in the strict sense (similar super-secondary arrangements and similar inter-repeat interactions), however, they have repetitive patterns of the secondary structures (or 2D TRs) (Fig. 6B). It means that among TRs predicted by *Tally* there might be a small fraction of proteins with 2D TRs (not 3D TRs). Finally, 8 of 22 misclassified cases are comprised of the structural core of the protein and do not have any repetitive secondary structure pattern.

3.2.2 TR-SS predicted as TR-SNS

In the benchmark set of 441 proteins with 3D TRs, there are 73 misclassified cases where 3D TRs are not detected in the sequence according to *Tally*.

Twenty-two of them have been found by HHrepID (Biegert and Söding, 2008). This program is sensitive at finding those hidden cases because it uses additional information from MSA of homologous proteins (Fig. 6C). The 51 remaining cases have been found by TRUST (Szklarczyk and Heringa, 2004), that infers homology information between the repeats through the concept of transitivity. Twenty-eight of these 51 TRs have only 2 repeat units which could explain why they got penalized in the scoring process. Indeed, TRs with only two repeats in structure and sequence are usually difficult to detect. Furthermore, analysis of the remaining 23 cases shows that some of them can be unveiled by *Tally* after their MSA optimization (Fig. 7). This suggests that some MSA generated by the TR predictors may require a post treatment in order to be optimized before being scored by *Tally*.

3.2.3 Performance of Tally in different types of TR-SS

Classification of known 3D TR structures uncovers a straightforward relationship between their architecture and the length of the repetitive units (Kajava, 2012). The 3D TRs can be subdivided into five major classes: class I – crystalline aggregates formed by regions with 1 or 2 residue long repeats; class II – fibrous structures stabilized by interchain interactions with 3–7 residue repeats; class III – structures with the repeats of 5–40 residues dominated by solenoid proteins; class IV 'closed' (not elongated) structures with the repeats of 30–60 residue long; and finally, class V – structures with typical size of repeats over 50 residues, which are large enough to fold independently into stable domains. The class V structures display a 'beads on a string' organization with 'beads' corresponding to globular domains. It was interesting to check how *Tally* performs in each class of these proteins. Supplementary Figure S2 shows the results of the analysis. Structures in classes III and IV are well recognized with only 11 and 9% of false positives respectively. At the same time, *Tally* was not able to detect 3D TRs in 31 and 19% of proteins from class II (fibrous structures) and V (bead-on-string structures), respectively. Class II contains many α-helical coiled coils, mainly identified by HHRepID, which are known to be difficult to detect in sequence by other methods due to the degenerated repeat motifs. Class V gathers structures with long repetitive units that fold into independent domains. Almost all of the 36 cases that are not identified in this class they have only 2 repeat units (29/36). As previously mentioned the 3D TRs with 2 repetitive units and the imperfect repeat motif represent a difficult case for the TR detection. The analysis also reveals that MSAs of some other non-detected TRs can be optimized (2/36) (Fig. 7). To solve this problem a post treatment of the MSAs consisted of removal of the terminal repeats with long arrays of gaps, can be apply prior to the scoring.

4 Conclusions

The survey of currently available computational programs for detection of TRs in protein sequences shows no best approach exists to cover the whole range of repeats (Kajava, 2012). This evokes the necessity to use a combination of several software products integrated into a pipeline. However, each of these programs uses its own scoring procedures to determine the boundary between repetitive and non-repetitive protein sequences. The concordance between sequence TRs, structure TRs and ultimately their function is of particular interest. Our survey showed that none of the existing scoring procedures that evaluate the presence of TRs in sequences are able to achieve an appropriate separation between genuine structural TRs and aperiodic structures (Richard and Kajava, 2015). This suggests that if we want to obtain a collection of structurally and functionally meaningful TRs for large scale analysis of proteomes, the TR scoring metrics need to be improved.

Here, we presented a scoring tool *Tally* that is able to fill this gap. An additional advantage of *Tally* is its universality because it requires only the MSA of the TR without prior information.

Therefore, it allows us to score any kind of TRs detected by different predictors and is adapted to large scale analysis by a Meta Repeat Finder. Moreover, it can also be used to reduce the false positive rate of the current embedded scoring systems, such TRUST or HHrepID when applied on their outputs. Available online, *Tally* can also be used to evaluate a MSA generated by users. Additionally, we provide a curated dataset of TR-SS and TR-SNS proteins that can be further used for benchmarking purposes.

Tally is developed specifically for proteins with 3D structures. This is the major case of TRs, however, some TRs are also located in the intrinsically disordered (or naturally unfolded) region. One can imagine that the evolutionary scenarios of TR regions folded in stable 3D structure and natively unfolded TRs are quite different (Richard and Kajava, 2015). The natively unfolded TRs do not have periodic structural constraints. In the absence of the 3D structure it is difficult to find a clear criterium for true and false TRs within unstructured regions, and as a result, this complicates the benchmark of existing TR scoring functions. Further studies are required to establish whether *Tally* or the other types of scoring approaches are the most appropriate for the natively unfolded TRs.

Acknowledgements

The authors thanks Frédéric de Lamotte, Eric Rivals, Daniel B. Roche, Etienne Villain for discussions; Phuong Do Viet for help with the construction of the dataset; Gaëtan Droc and Alexis Dereeper for help making the tool available online; We are grateful to the Genotoul bioinformatics platform Toulouse Midi-Pyrenees (Bioinfo Genotoul) for providing computing resources.

Funding

Ministère de l'Enseignement supérieur et de la Recherche (MESR) and Fondation pour la Recherche Médicale (FDT20150532625) grants to FDR; RA has a fellowship at the IBC (ANR 'investissement d'avenir en bioinformatique-projet-IBC'). COST Action BM1405 grant to AVK.

Conflict of Interest: none declared.

References

Andrade,M.A. *et al*. (2001) Protein repeats: structures, functions, and evolution. *J. Struct. Biol*., **134**, 117–131.

Biegert,A. and Söding,J. (2008) De novo identification of highly diverged protein repeats by probabilistic consistency. *Bioinf. Oxf. Engl*., **24**, 807–814.

Bucher,P. *et al*. (1996) A flexible motif search technique based on generalized profiles. *Comput. Chem*., **20**, 3–23.

Chen,X. and Ishwaran,H. (2012) Random forests for genomic data analysis. *Genomics*, **99**, 323–329.

Chothia,C. and Lesk,A.M. (1986) The relation between the divergence of sequence and structure in proteins. *Embo J*., **5**, 823.

Chou,P.Y. and Fasman,G.D. (1974) Prediction of protein conformation. *Biochemistry (Mosc.)*, **13**, 222–245.

Di Domenico,T. *et al*. (2014) RepeatsDB: a database of tandem repeat protein structures. *Nucleic Acids Res*., **42**, D352–D357.

Do Viet,P. *et al*. (2015) TAPO: a combined method for the identification of tandem repeats in protein structures. *FEBS Lett*., **589**, 2611–2619.

Edgar,R.C. *et al*. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*., **32**, 1792–1797.

Flach,P. (2012) *Machine Learning: The Art and Science of Algorithms that Make Sense of Data* Cambridge University Press, Cambridge.

Heringa,J. (1998) Detection of internal repeats: how common are they? *Curr. Opin. Struct. Biol*., **8**, 338–345.

Jorda,J. and Kajava,A.V. (2009) T-REKS: identification of Tandem REpeats in sequences with a K-meanS based algorithm. *Bioinf. Oxf. Engl*., **25**, 2632–2638.

Kajava,A.V. (2012) Tandem repeats in proteins: From sequence to structure. *J. Struct. Biol*., **179**, 279–288.

Kondo,H. *et al*. (2011) Ice-binding site of snow mold fungus antifreeze protein deviates from structural regularity and high conservation. *Proc. Natl. Acad. Sci. U. S. A*., **109**, 9360–9365.

Larkin,M.A. *et al*. (2007) Clustal W and Clustal X version 2.0. *Bioinformatics*, **23**, 2947–2948.

Liggett,W.H. and Sidransky,D. (1998) Role of the p16 tumor suppressor gene in cancer. *J. Clin. Oncol*., **16**, 1197–1206.

Li,W. and Godzik,A. (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, **22**, 1658–1659.

Luh,F.Y. *et al*. (1997) Structure of the cyclin-dependent kinase inhibitor p19Ink4d. *Nature*, **389**, 999–1003.

Marcotte,E.M. *et al*. (1999) A census of protein repeats. *J. Mol. Biol*., **293**, 151–160.

Matthews,B.W. (1975) Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta*, **405**, 442–451.

Mendoza,M.R. *et al*. (2013) RFMirTarget: predicting human MicroRNA target genes with a random forest classifier. *PLoS ONE*, **8**, e70153.

Morin,P.J. (1999) beta-catenin signaling and cancer. *BioEssays News Rev. Mol. Cell. Dev. Biol*., **21**, 1021–1030.

Newman,A.M. and Cooper,J.B. (2007) XSTREAM: a practical algorithm for identification and architecture modeling of tandem repeats in protein sequences. *BMC Bioinformatics*, **8**, 382.

Paladin,L. *et al*. (2015) Structural in silico dissection of the collagen V interactome to identify genotype-phenotype correlations in classic Ehlers–Danlos Syndrome (EDS). *FEBS Lett*., **589**, 3871–3878.

Pedregosa,F. *et al*. (2011) Scikit-learn: machine learning in python. *J. Mach. Learn. Res*, **12**, 2825–2830.

Remmert,M. *et al*. (2012) HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nat. Methods*, **9**, 173–175.

Richard,F.D. and Kajava,A.V. (2014) TRDistiller: A rapid filter for enrichment of sequence datasets with proteins containing tandem repeats. *J. Struct. Biol*., **186**, 386–391.

Richard,F.D. and Kajava,A.V. (2015) In search of the boundary between repetitive and non-repetitive protein sequences. *Biochem. Soc. Trans*., **43**, 807–811.

Schaper,E. *et al*. (2012) Repeat or not repeat?—statistical validation of tandem repeat prediction in genomic sequences. *Nucleic Acids Res*., **40**, 10005–10017.

Shimizu,T. *et al*. (2001) Active-site architecture of endopolygalacturonase I from *Stereum purpureum* revealed by crystal structures in native and ligand-bound forms at atomic resolution. *Biochemistry (Mosc.)*, **41**, 6651–6659.

Simeonova,I. *et al*. (2012) Fuzzy tandem repeats containing p53 response elements may define species-specific p53 target genes. *PLoS Genet*., **8**, e1002731.

Szklarczyk,R. and Heringa,J. (2004) Tracking repeats using significance and transitivity. *Bioinf. Oxf. Engl*., **20**, i311–i317.

Tonkin,M.L. *et al*. (2012) Structural and Functional Divergence of the Aldolase Fold in Toxoplasma gondii. *J. Mol. Biol*., **427**, 840–852.

Touw,W.G. *et al*. (2015) A series of PDB-related databanks for everyday needs. *Nucleic Acids Res*., **43**, D364–D368.

Valdar,W.S.J. (2002) Scoring residue conservation. *Proteins Struct. Funct. Bioinf*., **48**, 227–241.

Walsh,I. *et al*. (2015) Correct machine learning on protein sequences: a peer-reviewing perspective. *Brief. Bioinf*., doi: 10.1093/bib/bbv082.

Williams,R.W. *et al*. (1987) Secondary structure predictions and medium range interactions. *Biochim. Biophys. Acta BBA – Protein Struct. Mol. Enzymol*, **916**, 200–204.

Zhang,Y. (2005) TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res*., **33**, 2302–2309.