Tally-2.0: upgraded validator of tandem repeat detection in protein sequences

Vladimir Perovic, Jeremy Leclercq, Neven Sumonja, Francois Richard, Nevena Veljkovic, Andrey Kajava

Tally-2.0: upgraded validator of tandem repeat detection in protein sequences

Vladimir Perovic[1], Jeremy Leclercq[2], Neven Sumonja[1], Francois D. Richard[2,3], Nevena Veljkovic[1] and Andrey V. Kajava[2]

[1] Laboratory for Bioinformatics and Computational Chemistry, Institute of Nuclear Sciences VINCA University of Belgrade for Multidisciplinary Research, Institute of Nuclear Sciences VINCA, University of Belgrade, Belgrade, Serbia

[2]Centre de Recherche en Biologie cellulaire de Montpellier (CRBM), UMR 5237 CNRS, Université de Montpellier, Montpellier 34293, France

[3] Laboratory for Translational Breast Cancer Research, Department of Oncology, KU Leuven, 3000 Leuven, Belgium.

Abstract

Summary: Proteins containing tandem repeats (TRs) are abundant, frequently fold in elongated non-globular structures and perform vital functions. A number of computational tools have been developed to detect TRs in protein sequences. A blurred boundary between imperfect TR motifs and non-repetitive sequences gave rise to necessity to validate the detected TRs. Tally-2.0 is a scoring tool based on a machine learning approach, which allows to validate the results of TR detection. It was upgraded by using improved training datasets and additional machine learning features. Tally-2.0 performs at a level of 93% sensitivity, 83% specificity and an Area Under the Receiver Operating Characteristic Curve of 95%.

Availability and implementation: Tally-2.0 software is available, as a free web tool and as a standalone application published under Apache License 2.0, on the URL: https://bioinfo.crbm.cnrs.fr/index.php?route=tools&tool=27. It is supported on Linux. Source code is available upon request.

Contact: andrey.kajava@crbm.cnrs.fr

Supplementary information: Supplementary data are available at Bioinformatics online.

1. Introduction

Numerous studies demonstrate the fundamental functional importance of protein regions containing periodic sequences representing arrays of similar motifs that are directly adjacent to each other. The majority of proteins with these tandem repeats (TRs) in sequences have repetitive non-globular arrangements in their 3D structures (Kajava, 2012; Fraser and MacRae, 1973). Functions of these protein regions also frequently differ from the protein domains having aperiodic sequences folded in the globular structures. The TRs containing proteins predominantly serve as structural blocks (e.g. collagen, silk, keratin, proteins of epithelial tissues), as large hub proteins involved in protein–protein interactions (LRR or HEAT proteins), as core elements of multi-protein machineries and as proteins used like multivalent binders of ligands with periodic structures (Fraser and MacRae, 1973; Kobe and Kajava, 2001; Andrade and Bork, 1995)

The structural and functional differences of proteins with aperiodic and periodic sequences, points to the importance of bioinformatics tools that are able to distinguish between these two types of sequences. Most of the existing methods (Jorda and Kajava, 2009; Szklarczyk and Heringa, 2004; Biegert and Söding, 2008) can detect perfect TRs; however, in many cases, TRs are imperfect, contain a number of mutations accumulated during evolution and cannot be easily identified. In this situation, the 3D structure of proteins can be used as a benchmarking criterion for TR detection in sequences. The majority of proteins having TRs are built of repetitive 3D structural blocks and, the evolution cannot completely erase the repetitive patterns because some residues located in the equivalent positions of the repeats are critical for maintenance of the stable and functional structure. Previously, we developed a scoring tool called “Tally”, which is based on a machine learning (ML) approach and trained and evaluated on curated datasets 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) (Richard *et al.*, 2016). Tally achieved a better separation between sequences with structural TRs and sequences of aperiodic structures, than the other existing scoring procedures. In this work, we significantly improved this scoring tool by using additional ML features and enlargement of the curated benchmarking datasets. The dataset of “true” TRs was enriched in nearly perfect TRs allowing us to extend Tally application to the TRs of the natively unfolded regions.

2. Materials and methods

Datasets

Previously, we built a positive set of 441 "true" TRs found both in sequence and in structure and 141 'false' TRs only found in sequence but not in the structure (Richard *et al.*, 2016). Here, we improved these datasets by (1) increasing and equalizing numbers of TRs in the positive and negative datasets (553 and 525, correspondingly), (2) verifying and decreasing TR sequence redundancy in both datasets and (3) choosing TRs that allow a more equal representation in terms of their perfection, length and number of repeats. The TR of a given region is presented as multiple sequence alignment (MSA) of its repeats. For the TR identification and generation of MSAs, T-REKS (Jorda and Kajava, 2009), TRUST (Szklarczyk and Heringa, 2004) and HHrepID (Biegert and Söding, 2008) programs were used.

Machine learning algorithm and features

Previously, we generated 40 MSAs based ML feature (Richard *et al.*, 2016). In this work, we added 3 new features related to the number of gap openings in the MSA, and also a new family of 112 features, which are based on Fourier Transform and physico-chemical characteristics of amino acids. These spectral features are developed based on Informational Spectrum Method (ISM) (Veljkovic *et al.*, 2007) and are comprising of 4 groups: (1) two features based on amplitude values of first peaks in spectral representations of MSA, (2) eight features, which represent sum of signal/noise values on spectral peaks, (3) one noise based feature, and (4) three entropy based features, across 8 amino acid characteristics from AAIndex database (Nakai *et al.*, 1988) (see Supplementary data). In feature engineering process we have selected 55 from total of 155 original attributes for final model using sequential backward elimination (Saeys *et al.*, 2007) as a feature selection algorithm (see Supplementary data). The backward feature elimination was done by using H2O.ai platform (2018) and custom implementation in R language. The H2O.ai platform (2018) was used for cross validation process. The Tally-2.0 classifier was generated using Random Forest (Breiman, 2001) classification ML algorithm, as a method with the best prediction efficacy (see the comparison in Supplementary data).

3. Results

Tally-2.0 classifier was implemented in JAVA language using ML platform H2O.ai (2018). As an input, Tally-2.0 uses the list of TR regions presented as MSAs of their repeats. The calculation of MSA based features

is implemented in Python and of Spectral features in JAVA. The output lists Tally-2.0 score and several other known TR scores (Psim, entropy, p-value-phylo and parsimony (Richard and Kajava, 2015) allowing the users to validate the quality of the examined TRs. Tally-2.0 just like Tally (Richard *et al.*, 2016) has the best performance when we use Random Forest classifier (see Supplementary data), which indicates that the better results of the upgraded tool is due to the improved training datasets and additional ML features.

The evaluation of Tally-2.0, carried out on 10-fold cross-validation, showed 0.95 of Area Under the Receiver Operating Characteristic Curve (AUC) (Figure 1a). At a threshold of 0.45, established based on the maximization of F-score, Tally-2.0 performs at the level of 0.88 accuracy, 0.89 F-score, 83% specificity, while achieving a high value 93% of sensitivity. In addition, we compared Tally-2.0 to existing scoring methods as follows: Tally-2.0 scores was obtained with 10-fold cross-validation on the positive and negative training set, while the performance of the other scoring methods was evaluated by the direct calculation of the scores of the complete training set. Our comparative analysis showed that Tally-2.0 evaluates the separation between sequences with and without TRs better than the other scoring procedures (Figure 1).

Initially, Tally was developed to distinguish between protein structures with repetitive and non-repetitive architectures and, therefore, its dataset was enriched in MSAs that were close to the boundary between these two classes of proteins. As a result, Tally did not score well the MSAs which were far apart from this boundary (e.g. almost perfect repeats or MSAs from aperiodic random sequences) (Richard *et al.*, 2016). The updated dataset of “true” TRs used to build Tally-2.0 was enriched, on the one hand, in the perfect and almost perfect TRs and, on the other hand, in the random aperiodic sequences. It is also important to note that Tally input requires only sequence information. All this allowed us to cover the whole spectrum of MSAs and to extend application of Tally to the TRs of the natively unfolded (or intrinsically disordered) regions. Now, Tally 2.0 can be used in the large scale analyses as a uniform validator of TR detection. It is one of the most important application of our tool as at present each of TR detection programs use their own scoring measure. As a result, in the previous large scale surveys, the number of TR containing proteins in the proteomes varied significantly (between 14 to 30 %) (Marcotte *et al.*, 1999; Pellegrini, 2015) and the question about the total number of TRs in proteomes still stand unanswered.

Thus, the standalone version of Tally-2.0 is suitable for the validation of the large-scale analysis of TRs. In addition, web-based version of Tally-2.0 allows the users to validate imperfect TRs identified by them in the protein of their interest.

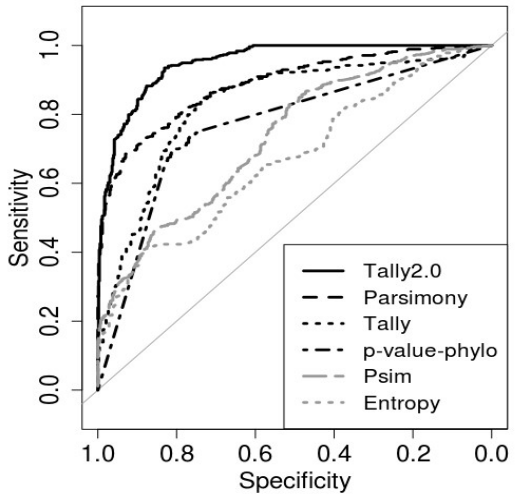

Figure 1

Comparative analysis of TR validators. For Tally-2.0, ROC curve has been obtained on the training set with 10-fold cross-validation, whereas for the other existing scoring methods we used the Tally-2.0 training dataset. Values of AUC in decreasing order are 0.95, 0.89, 0.83, 0.77, 0.73 and 0.67, respectively, for Tally2.0, Parsimony, Tally, p-value-phylo, Psim and Entropy scores.

Funding

Supported by the H2020-MSCA-RISE project REFRACT - GA No. 823886, by the National Institute of Allergy and Infectious Diseases (Research Grant 1R01AI12123701) and by grant No. 173001 (to V.P., N.S. and N.V.) from the Ministry of Education, Science and Technological Development, Republic of Serbia. This work was done within COST Action BM1405.

References

Andrade,M.A. and Bork,P. (1995) HEAT repeats in the Huntington's disease protein. *Nat. Genet.*, **11**, 115–116.

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

Breiman,L. (2001) Random forest. Machine learning Springer.

Fraser R.D.B. and MacRae,T.P. (1973) Conformation in Fibrous Proteins and Related Synthetic Polypeptides Elsevier.

Jorda,J. and Kajava,A.V. (2009) T-REKS: Identification of Tandem REpeats in sequences with a K-meanS based algorithm. *Bioinformatics*, **25**.

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

Kobe,B. and Kajava,A.V. (2001) The leucine-rich repeat as a protein recognition motif. *Curr. Opin. Struct. Biol.*, **11**.

Marcotte,E.M. *et al.* (1999) A census of protein repeats. *J. Mol. Biol.*, **293**, 151–60.

Nakai,K. *et al.* (1988) Cluster analysis of amino acid indices for prediction of protein structure and function. *Protein Eng.*, **2**, 93–100.

Pellegrini,M. (2015) Tandem Repeats in Proteins: Prediction Algorithms and Biological Role. *Front. Bioeng. Biotechnol.*, **3**, 143.

Richard,F.D. *et al.* (2016) Tally: a scoring tool for boundary determination between repetitive and non-repetitive protein sequences. *Bioinformatics*, **32**, 1952–1958.

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.

Saeys,Y. *et al.* (2007) A review of feature selection techniques in bioinformatics. *Bioinformatics*, **23**, 2507–2517.

Szklarczyk,R. and Heringa,J. (2004) Tracking repeats using significance and transitivity. *Bioinformatics*, **20 Suppl 1**, i311-7.

Veljkovic,V. *et al.* (2007) Application Of The EIIP/ISM Bioinformatics Concept in Development of New Drugs. *Curr. Med. Chem.*, **14**, 441–453.

Supplementary Data

Tally-2.0: upgraded validator of tandem repeat detection in protein sequences

Vladimir Perovic, Jeremy Leclercq, Neven Sumonja, Francois D. Richard, Nevena Veljkovic and Andrey V. Kajava

Features used for the classifier

In total, 155 original features were used for the classification. They consist of 40 Multiple Sequence Alignment (MSA) based features which were described in our paper on the previous version of Tally [Richard, Alves and Kajava, 2016], 3 new gap-related features (Table 1) and 112 new spectral features (Table 2).

Table 1. Gap-opening features related to the number of gap openings in the MSA of tandem repeats.

Name	**Short name**
Gap-open block measure	gapopen
Relative gap-open measure	rel_gapopen
Gap-open measure per amino acid	gapopen_peraa

Spectral features are developed based on Informational Spectrum Method [Veljkovic et al., 2007] where the amino acid sequence is first encoded into series of real numbers encoding each amino acid using its specific physico-chemical characteristic. This vector is in the second step transformed into Informational Spectrum (IS) using Fourier Transform. Family of 112 spectral features is generated using 14 measures defined on IS (Table 2) across 8 amino acid characteristics listed in Table 3.

Table 2. Measures used for calculation of spectral features.

Name	**Short name**	**Family**
Amplitude value of the first peak	AMP_1	First peak
Signal-to-noise ratio of the first peak	SN_1	First peak
Sum of signal/noise values of first two peaks	SPSN_2	Sum of signal/noise
Sum of signal/noise values of first three peaks	SPSN_3	Sum of signal/noise
Sum of signal/noise values of first four peaks	SPSN_4	Sum of signal/noise
Sum of signal/noise values of first five peaks	SPSN_5	Sum of signal/noise
Sum of signal/noise values of first six peaks	SPSN_6	Sum of signal/noise
Sum of signal/noise values of first seven peaks	SPSN_7	Sum of signal/noise
Sum of signal/noise values of first eight peaks	SPSN_8	Sum of signal/noise
Sum of signal/noise values of all peaks	SPSN_all	Sum of signal/noise
Average of amplitude values in IS	Noise	Noise
Entropy calculated on all amplitude values of the IS	Entropy	Entropy
Entropy calculated on amplitudes of all peaks in IS	All_peaks_entropy	Entropy
Entropy calculated on signal/noise values in IS	SN_entropy	Entropy

Table 3. List of physico-chemical characteristics of amino acids used to calculate Spectral based features.

Name	Short name	Reference
Electron-ion interaction potential	eiip	[Veljkovic et al., 2007]
B-values	bval	[Vihinen et al., 1994]
DisProt	disp	[Campen et al., 2008]
FoldUnfold	fu	[Galzitskaya et al., 2006]
Hydrophobicity	hph	[Glisic et al., 2016]
TOP-IDP	idp	[Campen et al., 2008]
Side-chain mass	mass	[Huang and Chen, 2013]
Net charge	netc	[Klein et al., 1984]

Importance of features (Figure 1) was calculated based on Random Forest (RF) model. The variable importance is determined calculating the relative influence of each feature: whereas the variable is selected for splitting and how much the squared error improved as a result during the tree building process.

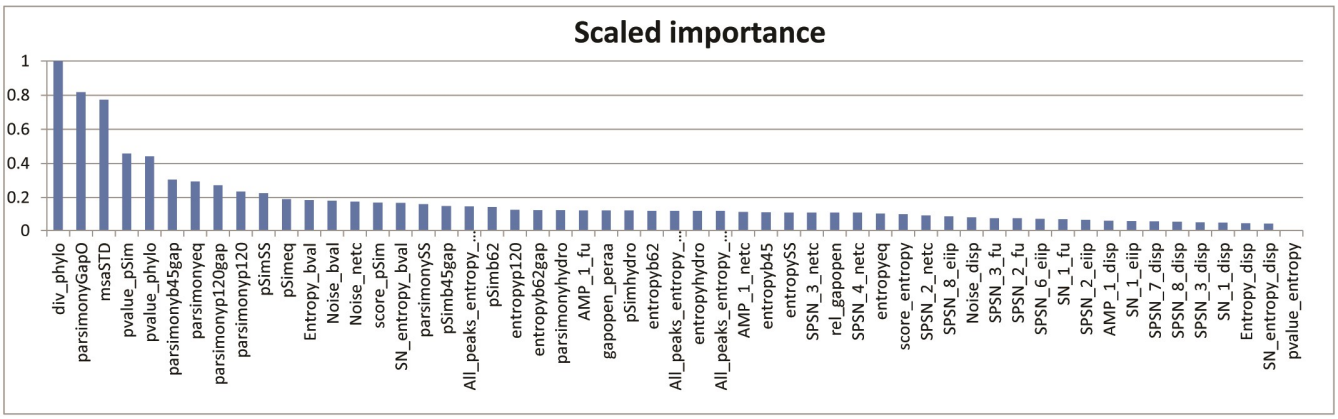

Figure 1.
Feature importance obtained when considering the 55 features of total 155 ones in the RF approach.

Machine learning algorithms comparison

Comparison of classifiers generated using different machine learning algorithms, Random Forest (RF) [Geurts et al., 2006], Gradient boosting machine (GBM) [Friedman and Jerome, 2001], Generalized linear model (GLM) [Breslow, 1996] and Deep learning (DEEPL) [Candel et al., 2015] was carried out using 10-fold cross-validation of the training set. Predictive performances were estimated by calculating accuracy (ACC), area under the ROC (receiver operating characteristic) curve (AUC), area under recall–precision plots (AUPRC), specificity, sensitivity, F1 score, precision and Matthews correlation coefficient (MCC) (Table 4).

Table 4. Comparison of the classifiers

	DRF	GBM	GLM	DEEPL

AUROC	0.950881	0.945248	0.927781	0.929534
AUPR	0.9506	0.948577	0.929398	0.92664
ACC	0.883117	0.87384	0.844156	0.845083
F1 score	0.891379	0.881119	0.856655	0.859546
Precision	0.853135	0.854237	0.812298	0.804724
Specificity	0.830153	0.835878	0.778626	0.763359
Recall	0.933213	0.909747	0.906137	0.922383
MCC	0.769035	0.748695	0.691971	0.696614

Runtime benchmark of Tally-2.0

Used dataset size: 202 065 MSAs.

Dataset statistics

Mean of repeat numbers in MSAs (n) :	4.43
Minimum repeat number (n_min) :	2
Maximum repeat number (n_max) :	100
Mean of repeat lengths in MSAs (l) :	78.52
Minimum repeat length (l_min) :	7
Maximum repeat length (l_max) :	1809

Computation time with 8 threads:
11 537 seconds (3,2047222 h)
Computer configuration:
Hard Disk Drive 7200 RPM
16 GO RAM – DDR3 - 1600MHz
Intel core I7-4770 – 3.40GHz – Cache 8192 KB
Linux Mint 19.2 "Tina" - Cinnamon (64-bit)

References

Richard, F. D., Alves, R. and Kajava, A. V. (2016). Tally: a scoring tool for boundary determination between repetitive and non-repetitive protein sequences, Bioinformatics, 32(13), pp. 1952–1958. doi: 10.1093/bioinformatics/btw118.

Veljkovic V, Veljkovic N, Este JA, Huther A, Dietrich U. Application of the EIIP/ISM bioinformatics concept in development of new drugs. Current medicinal chemistry. 2007;14(4):441-53.

Campen, A. et al. TOP-IDP-Scale: A New Amino Acid Scale Measuring Propensity for Intrinsic Disorder. Protein Pept. Lett. 15, 956–963 (2008).

Vihinen, M., Torkkila, E. & Riikonen, P. Accuracy of protein flexibility predictions. Proteins Struct. Funct. Bioinforma. 19, 141–149 (1994).

Galzitskaya, O. V., Garbuzynskiy, S. O. & Lobanov, M. Y. FoldUnfold: Web server for the prediction of disordered regions in protein chain. Bioinformatics 22, 2948–2949 (2006).

Glisic S, Cavanaugh DP, Chittur KK, Sencanski M, Perovic V, Bojic T. Common molecular mechanism of the hepatic lesion and the cardiac parasympathetic regulation in chronic hepatitis C infection: a critical role for the muscarinic receptor type 3. BMC bioinformatics. 2016;17(1):139.

Huang YF, Chen SY. Extracting physicochemical features to predict protein secondary structure. The Scientific World Journal. 2013; 2013.

Klein, P., Kanehisa, M. & DeLisi, C. Prediction of protein function from sequence properties. Discriminant analysis of a data base. Biochim. Biophys. Acta 787, 221–6 (1984).

Geurts P, Ernst D, Wehenkel L. Extremely randomized trees. Machine learning. 2006 Apr 1;63(1):3-42.

Friedman, Jerome H. “Greedy Function Approximation: A Gradient Boosting Machine.” Annals of Statistics (2001): 1189-1232.

Breslow NE. “Generalized Linear Models: Checking Assumptions and Strengthening Conclusions.” Statistica Applicata 8 (1996): 23-41.

Candel, Arno and Parmar, Viraj. “Deep Learning with H2O.” H2O.ai, Inc. (2015).