FEBS Letters xxx (2015) xxx–xxx

journal homepage: www.FEBSLetters.org

Review

TAPO: A combined method for the identification of tandem repeats in protein structures

Phuong Do Viet, Daniel B. Roche, Andrey V. Kajava *

Centre de Recherche de Biochimie Macromoléculaire, UMR 5237 CNRS, Université Montpellier, 1919, Route de Mende, 34293 Montpellier Cedex 5, France
Institut de Biologie Computationnelle, Université Montpellier, Bat. 5, 860, rue St Priest, 34095 Montpellier Cedex 5, France

ARTICLE INFO

Article history:
Received 29 June 2015
Revised 10 August 2015
Accepted 13 August 2015
Available online xxxx

Edited by Wilhelm Just

Keywords:
Tandem repeat
3D protein structure
Prediction of repetitive unit
Prediction pipeline
Webserver
Non-globular protein
Proteome

ABSTRACT

In recent years, there has been an emergence of new 3D structures of proteins containing tandem repeats (TRs), as a result of improved expression and crystallization strategies. Databases focused on structure classifications (PDB, SCOP, CATH) do not provide an easy solution for selection of these structures from PDB. Several approaches have been developed, but no best approach exists to identify the whole range of 3D TRs. Here we describe the TAndem PrOtein detector (TAPO) that uses periodicities of atomic coordinates and other types of structural representation, including strings generated by conformational alphabets, residue contact maps, and arrangements of vectors of secondary structure elements. The benchmarking shows the superior performance of TAPO over the existing programs. In accordance with our analysis of PDB using TAPO, 19% of proteins contain 3D TRs. This analysis allowed us to identify new families of 3D TRs, suggesting that TAPO can be used to regularly update the collection and classification of existing repetitive structures.

1. Introduction

Proteins can be broadly structurally classified as globular and non-globular. Non-globular proteins are mainly, disordered, membranous or repetitive. These repetitive proteins contain arrays of repeats that are adjacent to each other, called Tandem Repeats (TRs) (Fig. 1) [1–3]. Protein domains containing TRs are present in around one third of human proteins [4]. Recently developed methods for the identification of TRs in protein sequences [5–10] indicate that the number of TRs may be underestimated and we can expect an increasing number of TRs in proteins. These repetitive proteins are involved in a number of cellular activities, which include; maintenance of structural integrity (collagen and keratin); hub proteins involved in protein–protein interactions and as elements in multi cascade systems such as β-catenin and p16; ribonuclease inhibitors; catalytic activity (e.g. TIM-barrel proteins); phagocytosis and can be virulence factors [11–14]. As a result of TR proteins having a wide range of cellular activities they are implicated in a number of human diseases, which include cancers and neurodevelopmental disorders [13–16]. In addition, the growth of structural genomics initiatives, in combination with improvements in crystallographic and NMR techniques aimed at non-globular proteins, has resulted in an increase in structurally elucidated TR proteins deposited in the PDB [17]. The increase of available TR protein structures has necessitated the development of repeat protein classification schemes [11]. Structural repeats can be broadly divided into five classes mainly based on repeat length [11]; Class I – crystalline aggregates, such as polyalanine; Class II – fibrous structures such as collagen or α-helical coiled coils; Class III – elongated structures where the repetitive units require each other for structural stability such as solenoid proteins; Class IV – closed repetitive structures, which include TIM-barrels and β-propellers and Class V – bead on a string structures that include, for example, zinc finger proteins [11]. Recently, this classification was implemented in RepeatsDB database [18] where each of these classes have been further subdivided into several fine grained subclasses.

Despite this progress, however, the majority of bioinformatics approaches have been and remain to a large extent focused on globular proteins. In recent years, efforts have been made to

Author contributions: PDV developed the code for the pipeline, contributed text and figures for the manuscript. DBR supervised the work and drafted the manuscript and carried out final editing of the manuscript. AVK conceived the idea, supervised the work and drafted the manuscript and carried out final editing of the manuscript. All authors read and approved the final manuscript.

* Corresponding author at: Centre de Recherche de Biochimie Macromoléculaire, UMR 5237, Université Montpellier, 1919, Route de Mende, 34293 Montpellier Cedex 5, France. Fax: +33 (0)4 34 35 94 10.
E-mail address: andrey.kajava@crbm.cnrs.fr (A.V. Kajava).

http://dx.doi.org/10.1016/j.febslet.2015.08.025
0014-5793/

develop bioinformatics tools for the detection and analysis of repetitive elements in protein structures (3D TRs) such as feature-based learning methods RAPHAEL [19] and ConSole [20], a method for tiling structural space [21], a Fourier analysis method by Taylor et al. [22], wavelet transforms [23], signal analysis methods: DAVROS [24] and OPASS [25], methods that use conformational alphabets (ProStrip [26] and Swelfe [27]), and miscellaneous methods such as AnkPred [28], IRIS [29] and CE-Symm [30]. In the next section, we survey these approaches.

2. Survey of existing methods to identify tandem repeats in protein structure

2.1. Feature-based learning methods

RAPHAEL [19] is an algorithm that is especially good for the identification of tandem repeat protein structures of solenoid folds [31]. The method generates a periodicity profile for C_α atom coordinates (which is filtered by averaging the profile over 3 residue and 6 residue window). To avoid bias that is linked to the initial orientation of the structure, the protein is anchored to a reference point by random translation and rotation, which is repeated 200 times to produce more stable periodicity values. This and the other features based on structural periodicity and distance measurements are then combined using a Support Vector Machine (a machine learning approach) to differentiate between solenoid and non-solenoid proteins. Although RAPHAEL has been trained to detect solenoid protein structures, it is also able to detect the other structural classes of TRs. RAPHAEL is available as a webserver.

ConSole [20] is another feature based learning algorithm for the identification of solenoid proteins. ConSole uses image-processing, known as template-matching, to enable the differentiation between solenoids and non-solenoids. The template-matching process is applied to a contact map of the protein structure, with the resultant output feature scores (20 correlation coefficients) combined by applying a trained Support Vector Machine. ConSole is available as a webserver, with executable code available for download.

2.2. Signal analysis methods

A number of signal analysis based methods have been developed. Murray et al. [23] published a paper that studied how wavelet transforms could be used to detect and classify repeat motifs in structure, showing promising results on TIM-barrels and β-propeller structures [23]. This was one of the earliest *ab initio* methods, which, however, did not focus on the wide range of 3D TRs. In addition, the program based on this method is not publicly available, as the manuscript focused on the concept rather that the implementation. At the same time Taylor et al. [22] used Fourier analysis on a structural alignment scoring matrix comparing symmetry between two different substructures of the same protein [22]. The symmetrical structures appear as high scoring ridges, whose periods can be analyzed using the Fourier transform [22]. Again, this was one of the earliest studies in the field using Fourier transform, this analysis focused on proteins with internal symmetry and did not analyze the other TR-containing protein classes. This was followed by DAVROS [24], which utilizes the score matrix from a structural alignment program to aligned the protein under analysis on to itself (self-structural alignment), extracting the repetitive units from the matrix using a Fourier transform. DAVROS works well for repetitive proteins that do not contain large indels. Unfortunately, the source code for DAVROS is currently not available.

In addition, Parra et al. [21] developed a method that identifies "tiles" in the protein structure, basically potential repetitive units, using an exhaustive list of partial structural alignments, along with transformations of equivalent C_α atoms that maximize the superpositions. This produces non-overlapping "tiles" of the protein structure. Finally the method produces two scores in relation to the repetitive nature of the structure, a Tileability and Tile score. These score relates in a different but complimentary ways to the probability that a protein has a 3D TR. When the tile length is plotted versus the centre of each tile along the length of the protein structure, it produces a repetitive pattern for repeat proteins. Unfortunately, using these scores it is still difficult to classify automatically some repeat classes, such as TIM-barrels. The source code implementing this algorithm is currently not available.

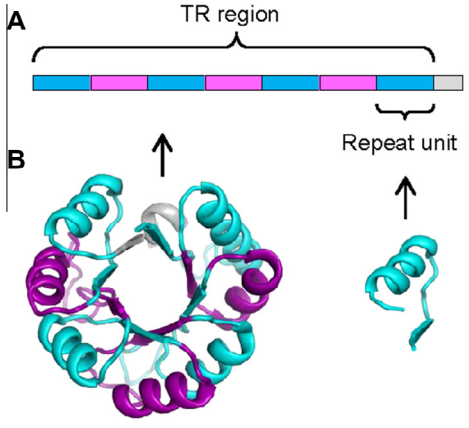

Fig. 1. A schematic representation of a protein containing TR region. (A) Protein sequence with TR regions highlighted and (B) 3D structure of the TR containing protein (in this case a TIM-barrel) along with the structure of one TR.

2.3. Conformational alphabet based methods

A number of algorithms have integrated conformational alphabet analysis for 3D TR detection, as repetitive protein structures have repetitive conformational sequences. These algorithms work best on repeat units longer than 20 residues. ProStrip [26], and Swelfe [32], basically uses protein backbone dihedral angles for every four consecutive C_α atoms, which are transformed into alphabet characters. This alphabet is used for fast scanning of proteins via dynamic programming, to determine if a protein structure is repetitive. The methods are available as webservers in addition to having source code available for download and use in-house.

2.4. Miscellaneous methods

Additionally, AnkPred [28], was recently developed to utilizes a graph-based approach, applying secondary structure feature based rules, for the identification of Ankyrin repeats in protein structures [28]. AnkPred is available as a webserver and downloadable for in-house use, however, it has not been designed to detect the wide range of repeat proteins currently classified.

Recently, a new method called CE-Symm [30] has been developed for the detection of internal symmetry in protein structures. This method has been developed specifically for a subclass of repeat proteins, which include TIM-barrels and β-propeller proteins. CE-Symm produces a score to determine internal symmetry, this score is an altered version of the TM-score [33], with the additional incorporation of symmetry order information. This study also focuses on the relationship of symmetrical proteins to enzyme functionality, symmetry around ligand binding sites in addition to tertiary and quaternary symmetry [30]. Unfortunately, if the protein structure

Please cite this article in press as: Do Viet, P., et al. TAPO: A combined method for the identification of tandem repeats in protein structures. FEBS Lett. (2015), http://dx.doi.org/10.1016/j.febslet.2015.08.025