Resources for repeat protein structure annotation: RepeatsDB and RepeatsDB-lite

Lisanna P(1)[†] , Hirsh L(1,2), Piovesan D(1) and Tosatto SCE(1,3)

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

[†] Email: lisanna.paladin@gmail.com

Motivation
Tandem repeats (TR) in proteins are ubiquitous in genomes and have been demonstrated to be of fundamental importance in many biological processes. They have several unique functions related to the development of organism complexity. TR proteins are characterized by a modular structure stabilized by a pattern of local interactions, which can be arranged in a wide variety of shapes providing functional diversity. Structural TR modules, called units, correspond to repeated segments in the sequence. Single units are associated to the protein function and can be used to classify and recognize different TR families. They are loosely conserved both at the DNA and amino acid level, therefore making their automatic recognition and classification a very hard task. The annotation of TRs at the unit level is essential because subtle differences in the structural conformation of the units give rise to large differences in shape and structural properties of the whole protein, ultimately determining their function and role. In recent year a number of new families has been discovered. The largest collection of TR proteins detected by structural features is provided by the RepeatsDB database. It relies on computational approaches and expert manual curation to detect TR in the Protein Data Bank (PDB) structures.
Methods
RepeatsDB was developed to fill the gap in TR protein annotation and provides the community with a high-quality resource of reliable datasets of repeat structures. Since is first release in 2014, RepeatsDB has been focused on improving the quality of unit annotation. In the new version (RepeatsDB 2.0), unit annotation has growth over an order of magnitude. This was possible by exploiting the repeat unit predictor in RepeatsDB-Lite web server. It uses an iterative structural search against a curated library of structural TR units to detect repetitive elements in PDB files. The TR unit library represents the conformational space and diversity of bona fide repeat units and covers all different TR classes. To increase predictor speed the unit library is hierarchically organized in three layers. The search starts from the reduced library and then propagates to other layers considering only related units. Several checks are introduced to minimize errors in the unit detection step and speed up the calculation. RepeatsDB-lite outperforms existing methods and can be applied to all types of TR proteins. The web interface allows to visualize similarity relationships between TR units at both the sequence and structure level. An all-against-all structure-based sequence similarity matrix is calculated and can be

used to get insights on the evolutionary relationship between the repeated units. The prediction can be manually refined by the user, visualizing the effects of the edits in real time. The web server allows an intuitive revision of the prediction and submission of reviewed entries to the database. The server represents a platform to harness community annotation efforts, which have been proven to be effective in RepeatsDB experience.

Results

The new release of RepeatsDB includes a new annotation pipeline producing extensive annotation for all entries. The pipeline is fully automated and allows the easy regular update of the database. The continuously growing number of RepeatsDB entries requires a continuous effort in the manual curation, to facilitate this process we designed RepeatsDB-lite, web server for the prediction and refinement of TR in protein structure. The revision process of the predictions guarantee high quality annotations. Currently about 60% of all entries are manually reviewed. All the entries are annotated with the unit position, for a total of ca. 6200 PDB chains containing ca. 6300 regions and more than 47000 units. New subclasses were identified and added to the structural classification scheme, currently including 4 classes and 23 subclasses. Moreover, a new classification level has been introduced on top of the existing scheme as an independent layer for sequence similarity relationships at 40%, 60% and 90% identity. RepeatsDB 2.0 includes completely redesigned web server and interface, to guarantee availability of data and better user experience in terms of database usability and look-and-feel. It features a new search engine for complex queries and a fully redesigned entry page for a better overview of structural data. It is now possible to compare unit positions, together with secondary structure, fold information and Pfam domains. Future work will concentrate on exploiting the repeat unit definitions to create profiles for use in detecting repeats from sequence for genome-scale analysis.