Journal of Structural Biology 179 (2012) 279–288

Contents lists available at SciVerse ScienceDirect

Journal of Structural Biology

journal homepage: www.elsevier.com/locate/yjsbi

Tandem repeats in proteins: From sequence to structure

Andrey V. Kajava *

Centre de Recherches de Biochimie Macromoléculaire, CNRS, Université Montpellier 1 et 2, 1919 Route de Mende, 34293 Montpellier, Cedex 5, France

ARTICLE INFO

Article history:
Available online 24 August 2011

Keywords:
Bioinformatics
Proteomes
Amino acid
Tandem repeat
3D structure

ABSTRACT

The bioinformatics analysis of proteins containing tandem repeats requires special computer programs and databases, since the conventional approaches predominantly developed for globular domains have limited success. Here, I survey bioinformatics tools which have been developed recently for identification and proteome-wide analysis of protein repeats. The last few years have also been marked by an emergence of new 3D structures of these proteins. Appraisal of the known structures and their classification uncovers a straightforward relationship between their architecture and the length of the repetitive units. This relationship and the repetitive character of structural folds suggest rules for better prediction of the 3D structures of such proteins. Furthermore, bioinformatics approaches combined with low resolution structural data, from biophysical techniques, especially, the recently emerged cryo-electron microscopy, lead to reliable prediction of the protein repeat structures and their mode of binding with partners within molecular complexes. This hybrid approach can actively be used for structural and functional annotations of proteomes.

1. Introduction

Dramatic growth of genomic data presents new challenges for scientists: making sense of millions of protein sequences requires systematic approaches and information about their 3D structure as well as about their evolutionary and functional relationships. The majority of protein sequences are aperiodic and usually have globular 3D structures carrying a number of various functions. The foremost efforts of researchers were devoted to these types of proteins and, as a result, significant progress has been made in the development of bioinformatics tools for their analysis. However, proteins also contain a large portion of periodic sequences representing arrays of repeats that are directly adjacent to each other (Heringa, 1998). These tandem repeats are considerably diverse, ranging from the repetition of a single amino acid to domains of 100 or more residues. They are ubiquitous in genomes and occur in at least 14% of all proteins (Marcotte et al., 1999). Furthermore, they are present in almost every third human protein and even in every second protein from *Plasmodium falciparum* or *Dictyostelium discoideum* (Pellegrini et al., 1999; Jorda and Kajava, 2010). The tandem repeat regions are highly polymorphic compared to the background rate of point mutations (Buard and Vergnaud, 1994; Ellegren, 2000). The two main mechanisms underlying this hypermutability are: (i) replication slippage within DNA microsatellite regions (repeat is less than 10 nucleotides) and (ii) recombination events for longer minisatellite and satellite regions.

Over the last decade, numerous studies demonstrated the fundamental functional importance of such tandem repeats and their involvement in human diseases. A number of evidences has also been gathered about the high incidence of tandem repeats in the sequences of virulence factors of pathogenic agents, toxins and allergens (Kajava et al., 2006). Genetic instability of these regions can allow a rapid response to environmental changes and thus can lead to emerging infection threats. Furthermore, tandem repeats frequently occur in amyloidogenic and other disease-related sequences (Baxa et al., 2006; Nelson and Eisenberg, 2006). This implies that this class of sequences may have a broader role in human diseases than was previously recognized.

Thus, tandem repeat regions are abundant in proteomes and are related to major health threats of the modern society. Along this line, the discovery of these domains, understanding of their sequence–structure–function relationship and mechanisms of their evolution promise to be a fertile direction for research leading to the identification of targets for new medicines and vaccines. However, the conventional bioinformatics approaches for annotation of proteomes that are developed for globular domains have limited success when applied to the tandem repeat regions. They require special computer programs and databases. Here, I survey available bioinformatics tools for analysis of protein repeats with emphasis on the sequences, 3D structures, sequence–structure relationship as well as highlighting successful strategies for the prediction of the protein structure.

* Fax: +33 4 34359599.
E-mail address: andrey.kajava@crbm.cnrs.fr

1047-8477/$ - see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.jsb.2011.08.009

2. Identification of tandem repeats in protein sequences

The growth of proteomic data has led to increasing efforts to develop methods for protein repeat recognition. Protein tandem repeats are frequently not perfect, containing a number of mutations (substitutions, insertions, deletions) accumulated during evolution, and some of them cannot be easily identified (Fig. 1). To solve this problem, over the last few years, several improved algorithms and software have been developed. They can be subdivided into five general types of methods (Table 1). The first type finds periodicities in the amino acid sequences by Fourier transform analysis (McLachlan and Stewart, 1976; Coward and Drablos, 1998; Gruber et al., 2005). In contrast to some other methods, this approach does not rely on *a priori* knowledge about putative repeats, representing a so called *de novo* or *ab initio* method. It is specialized for detection of long arrays of tandem repeats without insertions and deletions. These types of regions are frequently found in fibrous proteins such as collagen or α-helical coiled coil proteins. The second type of programs, such as XSTREAM (Newman and Cooper, 2007) or T-REKS (Jorda and Kajava, 2009), are based on short string extension algorithms. These methods can identify tandem repeats with indels and are especially good for *ab initio* detection of relatively short (less than 15–20 residues) repeats. They have a linear time complexity and, therefore, are rapid and well adapted for a large-scale search of protein repeats. The third type of methods detects repeats by comparing the protein sequence to itself with sequence–sequence alignment algorithms. Examples of the existing web-servers are RADAR and TRUST (Heger and Holm, 2000; Szklarczyk and Heringa, 2004). These programs are especially efficient for *ab initio* detection of arrays of long repeats (repetitive units of more than 15 residues), however, they frequently fail to identify short repeats. In addition, the sequence self-alignment algorithms with a time complexity of $O(n^2)$ (where n is the length of sequence), are relatively slow, and, do not suit large scale analysis well. The fourth type of approaches utilizes sets of *a priori* generated alignments of repeats which are used to construct Hidden Markov Models (HHMs) or sequence profiles (Gribskov et al., 1987; Bucher et al., 1996). The profiles or HMMs from these sets are compared one by one to the query sequence in search of the best and multiple hits to a repeat profile. One profile can span several repeats thus increasing the selectivity of the search compared to a single repeat (Kajava, 1998). The power of these methods depends on the quality of the sequence alignments used for construction of HMMs or profiles as well as on the completeness of the profile or HMM libraries. Currently available web servers with the tandem repeat HMMs or profiles are Pfam (Sonnhammer et al., 1998), SMART (Schultz et al., 1998), REP (Andrade et al., 2000), TPRpred (Karpenahalli et al., 2007), PROSITE (Hofmann et al., 1999) and BiSMM server (Kajava and Steven, 2006b). This approach is one of the best in detection of long and strongly imperfect tandem repeats, however, it requires *a priori* generated alignments of putative repeats and, therefore, is not suitable for automated *ab initio* large scale analysis. Finally, the fifth type of methods relies on HMM–HMM or profile–profile comparisons used, for example, in the HHrepID server (Biegert and Soding, 2008) for *ab initio* detection of tandem repeats. It constructs a HMM from a multiple alignment of proteins that are homologous to the analyzed one, and search for sub-optimal alignments of this HMM with itself. This approach can be very sensitive allowing efficient *ab initio* detection of highly divergent "covert" tandem repeats (Biegert and Soding, 2008; Kippert and Gerloff, 2009). This fifth type also includes methods of sequence profile comparison against discrete Fourier transform or stationary wavelet packet transform of sequences implemented in the programs REPETITA and WAVELET (Marsella et al., 2009; Vo et al., 2010). In contrast to *ab initio* method of HHrepID, these programs use *a priori* knowledge about repetitive patterns and are devoted to special repeats associated with solenoid protein structures. It is worth mentioning that the most sensitive methods for *ab initio* identification of "covert" tandem repeats are relatively slow and inappropriate for automated large scale analysis. The improvement of the algorithm rapidity presents one of the next challenges to the researchers because the genome and proteome data are growing dramatically-even faster than the power of computers.

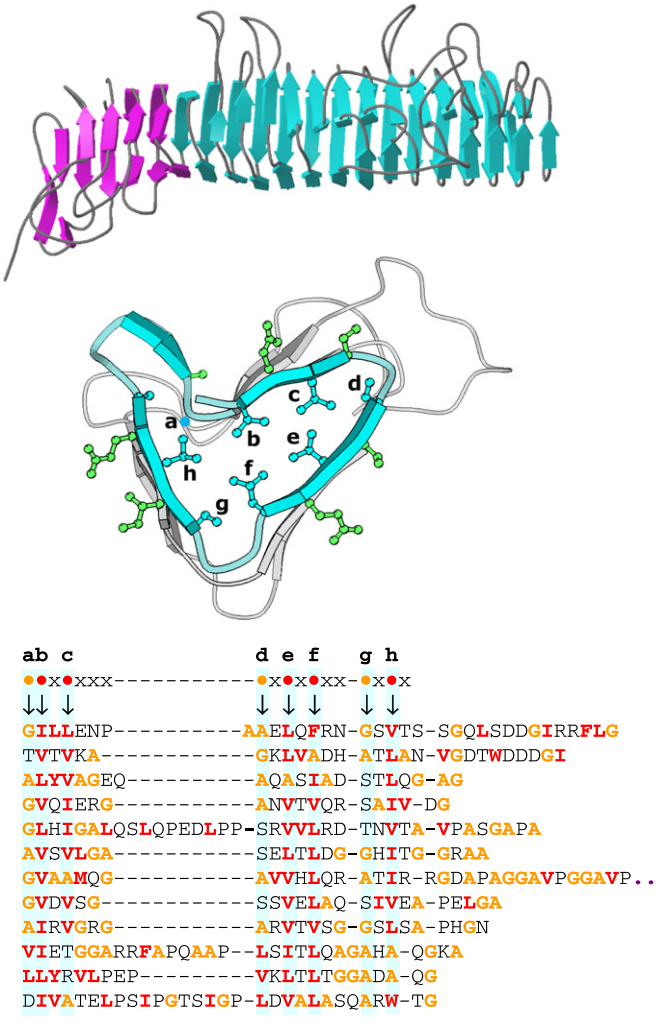

Fig.1. Pertactin from *Bordetella pertussis* is an example of a protein with highly degenerate repeating sequence that was uncovered only after the resolution of its 3D structure. (Top) Ribbon representation of the molecule with β-solenoid structure and its C-terminal capping domain shown in blue and magenta respectively. (Middle) Axial view of the β-solenoid coils. Letters a–h denote internal residues that are essential for the structural integrity of the fold. (Bottom) Alignment of the repeats. Arrows indicate a–h positions (highlighted by blue) preferentially occupied by bulky apolar residues (in red) or small apolar residues (orange).

Thus, over the last years, a number of efficient approaches for identification of protein repeats have been developed. Our survey of different available computer programs shows that depending on the size and character of the repeats some of them are performing better than others, but no best approach exists to cover the whole range of repeats. Therefore, today, the best way is to use a combination of several available software products to identify the repeats.