Chapter 46
A Web Database IR-PDB for Sequence Repeats of Proteins in the Protein Data Bank

Selvaraj Samuel
Bharathidasan University, India

Mary Rajathei
Bharathidasan University, India

ABSTRACT

Amino acid repeats play significant roles in the evolution of structure and function of many large proteins. Analysis of internal repeats of protein with known structure helps to understand the importance of repeats of the protein. A database IR-PDB for repeats in sequence of the proteins in the PDB has been developed for the analysis of impact of repeats in proteins. Using the state of the art repeat detection method RADAR, internal repeats in 148202 sequences out of 285714 sequences belonging to 115031 PDB structures were detected. The identified sequence repeats were annotated with secondary structural information with a view to analyze the structural consequence and conservation of the repeats. The tertiary structure of the repeats and their functional involvements can be found out through web links to PDB, PDBsum and Pfam. IR-PDB is systematically annotated for the proteins in the PDB with sequence repeats and their structure with the possibility to access the dataset interactively through web services.

INTRODUCTION

A large portion of proteins contain repeated segments of amino acids that often correspond to structural and functional units. The percentage of repeats containing proteins grows with the complexity of the organism which suggests that internal duplication is an important mechanism for the evolution of multicellular organisms. The repeat length varies considerably from few amino acids of shorter repeats (<50 in length) to larger span of domain repeats that can be present as repeat pair of single repeat to multiple number of repeats. Based on the distance between adjacent units, they are classified as tandem repeats of continuously distributed and non-tandem repeats of sequentially interspersed. It has been observed

DOI: 10.4018/978-1-5225-8903-7.ch046

that single amino acid/homopeptide repeats (Jorda & Kajava, 2010), oligopeptide (2-20 amino acids in length) (Fraser & MacRae, 1973) and greater than >20 amino acid repeats are involved in various diseases like neurodegenerative disorders, cancer, muscular dystrophy, and others (Djian, 1998; Peruz,1999; Orr et al., 2007; Burchel et al., 2006). It has been suggested that array of repeats provide regular spatial and functional groups which are useful for structural packing or for interactions with target molecules (Katti et al., 2000). Further, the involvement of different repeat types of length less than 60 such as tetratricopeptide, leucine-rich repeats, ankyrin and armadillo/heat repeats (Fraser & MacRae, 1973; Kobe & Kajava, 2001; Grover & Barford, 1999; Yoder et al., 1993) in various structures and functions of the proteins has been highlighted (Andrade et al., 2001). Analysis of repeat pairs of length >50 at the structure of the protein has shown that most of the sequence repeats adopt similar fold in spite of divergence and are involved in the function of the protein (Mary & Selvaraj, 2013). Study on conservation of tertiary structure between repeats in functional units (domains) of protein using structure based parameters suggests that equivalent residues in the repeated segments share similar tertiary environment for adopting similar fold (Mary, Saravanan, & Selvaraj, 2015).

A number of servers are available to detect sequence repeats in proteins, based on different algorithms. Web servers such as XSTREAM (Newman & Cooper, 2007), T-REKS (Jorda & Kajava, 2009) are based on short string extension algorithms which can identify tandem repeats with insertions and deletions of relatively short (less than 15-20 residues) repeats. The RADAR and TRUST web servers (Heger & Holm, 2000; Szklarczyk & Heringa, 2004) are efficient for the detection of long repeats (repetitive units of more than 15 residues) by comparing a protein sequence to itself. On the other hand, the TPRpred tool (Karpenahalli et al., 2007) and REP method (Andrade et al., 2000) use a priori generated alignments to construct Hidden Markov Models (HMMs) or sequence profiles (Bucher et al., 1996; Gribskov et al., 1987) to detect repeats. The profiles or HMMs from these sets are compared one by one to the query sequence in search of the best and multiple hits. Finally, HHrepID (Biegert & Soding, 2008) is a method that relies on both HMM-HMM or profile-profile comparison for 'ab initio' detection of tandem repeats.

Several databases for sequence repeats of proteins have been developed such as Homopeptide Repeats database for single amino acid repeats of the proteins from GenPept database of NCBI (Faux et al., 2005), COPASAAR for single amino acid repeats for all proteomes in EBI (Depledge & Dalby, 2005) and Tandem Repeats in a Protein Sequence (TRIPS) database for single amino acid as well as oligopeptide repeats from protein sequences in the SWISS-PROT (Katti et al., 2000). The general and well-annotated sequence databases such as UniProt (Uniprot Consortium, 2015), Pfam (Finn et al., 2016), SMART (Letunic et al., 2015) contain different tandem repeat types of tetratricopeptide, leucine-rich repeats, ankyrin and armadillo/heat, whereas the ProRepeat database contains both perfect and approximate tandem repeats in the sequences from UniProt, together with 85 completely sequenced eukaryotic proteomes in the RefSeq collection (Luo et al., 2012). Most of the available databases for sequence repeats contain only shorter tandem repeats of the proteins. However, proteins contain tandem and non-tandem repeats with varying lengths. Further, sequence repeats of proteins with known structure would help analyze the importance of repeats for the structure and function of the proteins. It also provides a unique opportunity to analyze repeats in different aspects such as conservation of overall fold of the duplicated region as well as changes in the structures due to the accumulation of mutations in the repeats which will be useful for protein design and engineering.

The rapid increase of proteins with known structures available in the Protein Data Bank (PDB) (Berman et al., 2007) and the lack of a large scale well-annotated repository for repeats in the sequences of proteins in the PDB have motivated us to construct a database IR-PDB (Internal Repeats of Proteins in