

DRIPS Database of Repeats in Protein Sequences

Rima Kumari
rima_kumari@research.iiit.ac.in

Nita Parekh
nita@iiit.ac.in

Center for Computational Natural Sciences and Bioinformatics
International Institute of Information Technology, Hyderabad, INDIA

Abstract

It has been observed that a large number of protein sequences in the SwissProt database contain repetitive regions, *viz.*, tandem peptide repeats, multiple copies of motifs/profiles, periodic occurrences of single amino acids, runs of single amino acids and multiple copies of domain which may or may not be tandemly repeated. Internal repetition affords a protein enhanced evolutionary prospects and provides opportunities for the organism to expand its repertoire of cellular functions, such as protein transport, protein-complex assembly, and protein regulation using pre-existing genetic material. We have developed a database of periodic repeats, DRIPS, which can be queried at the kingdom, proteome or individual protein level, allowing functional and evolutionary analysis of repeat patterns across organisms. (<http://ccnsb.iiit.ac.in/nita/PEPPER/PTRDB/>)

Keywords: protein repeats, tandem peptide repeats, single amino acid repeats

1 Introduction

Repeats are ubiquitous in protein sequences and occur in various forms *viz.*, periodic peptide repeats, multiple copies of motifs, profiles or domains which may or may not be tandemly repeated. These are thought to arise via intragenic duplication and recombination events. Selective advantage of multiple repeats result in these mutations being fixed in the population. Even though the ability to generate repeats appears to be a general phenomenon of all phyla, repeats are more common in eukaryotic organisms than in prokaryotic ones and in metazoans more than in the rest of the eukaryotes. This may possibly be due to the increased functionality provided by repeat assemblies in higher organisms, such as protein transport, protein-complex assembly, and protein regulation using pre-existing genetic material. Here we discuss the distribution of periodic repeats across kingdoms and discuss how the database can be useful for analyzing their evolution and functional role in proteins.

2 Method and Results

About the Database: A Database of Repeats in Protein Sequences (DRIPS) has been built on SwissProt database (ver 51.1) containing 3,59,542 proteins by self-comparison of sequences, using the in-house tool, PEPPER. A Perl script was used to extract functional information and cross-references to other databases from SwissProt files and the database built in Mysql. It consists of 29,812 proteins containing (exact and approximate) periodic repeats, *viz.* tandem peptide repeats

(TPRs), periodic occurrences of single amino acids (POSAAs) and single amino acid repeats (SAARs). The web-interface to DRIPS provides three types of query search (Fig. 1):

- Text search, e.g., Keyword Search or Organism Name
- Search by SwissProt Id, PDB Id, Repeat Length and Repeat Pattern.
- By combination of the above queries using Boolean operators (AND, OR, NOT).

The query search results a summary page (Fig. 2 (a)) listing the proteins matching the search criteria with a brief description and a hyperlink for details for each output. The details link opens a new page providing additional information about the repeat pattern, its alignment and cross-references to various databases such as, Family information (hyperlinked to Pfam) and description of associated disease with OMIM ID (if available).

Statistics of the Database: DRIPS consists of 6,760 proteins containing 9,640 TPRs; 11,005 proteins containing 12,888 SAARs and 12,047 proteins containing 19,219 POSAAs, corresponding to about 8.3% of proteins in SwissProt. However, only 1,071 sequences are annotated in SwissProt as tandem repeats and 6,386 proteins with compositional bias (i.e., SAARs), and no annotation is provided for POSAAs. About 75% proteins in DRIPS belong to eukaryotes. To check if this high abundance of repeats in eukaryotic proteins is due to longer sequence lengths, we analyzed the percentage of proteins containing TPRs by grouping proteins superkingdom-wise in bins of size 200. We observe that even for lengths < 600, eukaryotes have 2 – 3 times more number of repeats compared to Bacteria and Archaea, To understand the role of high abundance of repeats in eukaryotes, we looked into the protein classes with high incidence of repeats in eukaryotes and found that zinc finger proteins are most abundant (~ 15%) followed by collagen, antigen, chaperonin, keratin and prion (~ 6% each). These protein classes are not present in prokaryotes suggesting these to be specific to eukaryotes.

Functional Analysis: To analyze the functional role of repeats and their conservation across organisms, we first identified few most frequently occurring patterns in DRIPS of varying lengths. All proteins containing a particular pattern were subjected to PSI-BLAST to identify true homologs. The homologous proteins were then multiply aligned using CLUSTALW to analyze the loss/gain or conservation of copy number of repeats. For e.g., the heptapeptide, YSPTSPS, is found to occur in a large number of proteins: Q8SSC4, A5DCV3, Q75A34, P35084, P35074, P18616, P16356, and P04050. On multiple alignment it is observed that the copy number of this pattern is not conserved but increases from Fungi (14) to yeast (26), drosophila (32), C. elegans (41) to human & mouse (52). An OMIM link provided to repeats can be used to identify disease association, if reported. Thus, using this database, model organisms having homologs to human protein repeats can be identified and further investigated in the laboratory for functional analysis and disease association of repeats in protein sequences.

Database of Repeats In Protein Sequences (DRIPS)

[Home](#) | [Help](#) | [Pepper Tool](#) | [Contact Us](#)

Search Database

The Database can be searched with various options given using Boolean operators AND, OR, NOT

e.g. - search text[attribute]

e.g. P02817[SPID] AND QPL[PAT] OR NOT Bos taurus[ORG]

Search Option	Attributes
Swissprot Id	SPID
Organism Name	ORG
PDB ID	PDBID
Repeat Pattern	PAT
Repeat Length	RLEN
Keyword	KW
Copy Number	CN

Figure 1: Snapshot of DRIPS front page.

(a) Database of Repeats In Protein Sequences (DRIPS)

[Home](#) | [Help](#) | [Pepper Tool](#) | [Contact Us](#)

SUMMARY PAGE

P02817
 SEARCH QUERY: P02817[SPID]
 TOTAL NUMBER OF HITS: 3

SWISS PROT ID: [P02817](#)
 ORGANISM NAME: Bos taurus (Bovine)
 PROTEIN NAME: Amelogenin, X isoform precursor (Class I amelogenin)
 KEYWORDS: Alternative splicing, Biosynthesis, Direct protein sequencing, Extracellular matrix, Phosphoprotein, Repeat, Secreted, Signal
 PATTERN: P

SWISS PROT ID: [P02817](#)
 ORGANISM NAME: Bos taurus (Bovine)
 PROTEIN NAME: Amelogenin, X isoform precursor (Class I amelogenin)
 KEYWORDS: Alternative splicing, Biosynthesis, Direct protein sequencing, Extracellular matrix, Phosphoprotein, Repeat, Secreted, Signal
 PATTERN: Q

SWISS PROT ID: [P02817](#)
 ORGANISM NAME: Bos taurus (Bovine)
 PROTEIN NAME: Amelogenin, X isoform precursor (Class I amelogenin)
 KEYWORDS: Alternative splicing, Biosynthesis, Direct protein sequencing, Extracellular matrix, Phosphoprotein, Repeat, Secreted, Signal
 PATTERN: QPL

[Details](#) | [Submit a Page](#)
 1

(b) Database of Repeats In Protein Sequences (DRIPS)

[Home](#) | [Help](#) | [Pepper Tool](#) | [Contact Us](#)

DETAILS

SWISS PROT: [P02817](#)
 ORGANISM NAME: Bos taurus (Bovine)
 TAXONOMY ID: [9211](#)
 GENE NAME: AMELX
 PROTEIN NAME: Amelogenin, X isoform precursor (Class I amelogenin)
 FUNCTION: Plays a role in the homeostasis of teeth. Seems to regulate the formation of crystallites during the secretory stages of tooth enamel development.
 DISEASE: ---
 PFAM: [PF02246](#)
 PDB ID: NONE
 PATTERN: REPEAT LENGTH: QPL, 3
 COPY NUMBER: 5,300, 133,148
 START-END: ---
 ALIGNMENT:

```

...IQP-QPHQPLQPHQPLQPMQPMQPLQPLQPLQ-QPPY
QPLQPLQPLQPLQPLQPLQLQLQLQPLQPLQPLQ-P-
H
L
    
```

Figure 2: Output of DRIPS (a) summary page, (b) details about a repeat