DE GRUYTER Journal of Integrative Bioinformatics 2020; 17(4): 20200024

David Mary Rajathei, Subbiah Parthasarathy and Samuel Selvaraj*

HPREP: a comprehensive database for human proteome repeats

https://doi.org/10.1515/jib-2020-0024
Received June 8, 2020; accepted September 17, 2020; published online November 3, 2020

Abstract: Amino acid repeats are found to play important roles in both structures and functions of the proteins. These are commonly found in all kingdoms of life, especially in eukaryotes and a larger fraction of human proteins composed of repeats. Further, the abnormal expansions of shorter repeats cause various diseases to humans. Therefore, the analysis of repeats of the entire human proteome along with functional, mutational and disease information would help to better understand their roles in proteins. To fulfill this need, we developed a web database HPREP (http://bioinfo.bdu.ac.in/hprep) for human proteome repeats using Perl and HTML programming. We identified different categories of well-characterized repeats and domain repeats that are present in the human proteome of UniProtKB/Swiss-Prot by using in-house Perl programming and novel repeats by using the repeat detection T-REKS tool as well as XSTREAM web server. Further, these proteins are annotated with functional, mutational and disease information and grouped according to specific repeat types. The developed database enables the users to search by specific repeat type in order to understand their involvement in proteins. Thus, the HPREP database is expected to be a useful resource to gain better insight regarding the different repeats in human proteome and their biological roles.

Keywords: bioinformatics; database; disease; function; human proteome; repeats.

1 Introduction

Amino acid repeats that are commonly found in all kingdoms of life have played essential roles on both structures and functions of the proteins [1]. The different well-characterized repeats such as Leucine rich repeats (LRR), Ankyrin (ANK) and Armadillo etc., with regard to their structures and functions of the proteins have been extensively analysed [2–5]. The roles played by the domain repeats of immunoglobulin, human matrix metalloproteinase and zinc finger type proteins in protein–protein interaction as well as binding to DNA or RNA have been observed [6–8]. Several web servers and tools such as RADAR, TRUST, XSTREAM, T-REKS, and PTRStalker etc., that use different methods of sub-optimal alignment, short seed expansion, *K*-means clustering and normalized BLOSUM-weighted edit distance [9–13] for repeats detection have been developed. Further, the repeats in the sequences of Protein Data Bank (PDB) [14] and UniProtKB/Swiss-Prot [15] identified by using RADAR have been analyzed at the structural and functional level. Several databases for amino acid repeats from different set of protein sequences were constructed for large scale analysis. RepSeq [16] is a database for repeats of lower eukaryotic pathogens obtained by searching identical shorter amino acids, PTRStalkerDB for repeats of SwissProt identified by using PTRStalker algorithm [13] and ProRepeat database [17] for repeats in UniProt as well as in some eukaryotic proteomes of RefSeq collection obtained based on suffix tree algorithm. The PRDB includes repeats found in the sequences of (i) NR (non-redundant) data bank of NCBI (ii) PDB and (iii) Swiss-Prot obtained by using T-REKS program [18]. The IR-PDB database for different

---

***Corresponding author: Samuel Selvaraj**, Department of Bioinformatics, School of Life Sciences, Bharathidasan University, Tiruchirappalli 620 024, Tamil Nadu, India, E-mail: selvarajsamuel@gmail.com
David Mary Rajathei and Subbiah Parthasarathy, Department of Bioinformatics, School of Life Sciences, Bharathidasan University, Tiruchirappalli 620 024, India, E-mail: rajathei@yahoo.com (D. Mary Rajathei), bdupartha@gmail.com (S. Parthasarathy). https://orcid.org/0000-0003-0089-0595 (D. Mary Rajathei)

repeats patterns of tandem repeats, non-tandem repeats, shorter repeats and long repeats identified in the sequences of PDB by using RADAR has been developed [19]. The various repeat detection algorithms, web servers, tools, databases and their online availability have been listed out [20]. It has been observed that protein repeats are not only carrying out important biological functions but also related to several diseases. The high incidence of tandem repeats in the sequences of virulence factors of pathogenic agents, toxins, and allergens and in other disease-related sequences have been found out [21–23].

Earlier, the Human Genome Project (HGP) has derived a draft map of complete human proteome with approximately 20,000 protein-coding genes. These proteins are expertly curated and annotated with amino acid sequence, protein name, protein family, repeats, domain, function, mutation and disease in UniProtKB/Swiss-Prot [24]. The normal functions of single amino acid repeats [25] and their abnormal expansion for several human diseases [26–29] have been studied. Further, the PolyQ 2.0 database for polyglutamine (polyQ) repeats of human proteins with functional, domain and single point mutation information has been developed [30]. It was also observed that 15–20% of the human proteins have contained repeats of size longer than 5 [31, 32]. However, there is no exclusive repository for human proteins containing different categories of repeats of longer in lengths. Towards this goal, *a* web database HPREP that consists of different categories of well-characterized, domain and novel repeats of human proteins provided with functional, mutational and diseases information was developed.

2 Material and methods

The work flow of generation and accessing of the HPREP database is shown as a flowchart (Figure 1). First we detect the well-characterized repeats and domain repeats that are present in the human protein sequences of UniProtKB/Swiss-Prot. Then, the proteins with no repeats assignment are analysed for novel repeats by using repeat detection T-REKS tool as well as XSTREAM web server. Further, the functions, mutations and diseases of the repeat proteins are assigned and developed as a HPREP database. The database can be accessed either by giving UniProtID/Protein name or by selecting specific well-characterised repeat name, domain name and novel repeat containing protein. The database was implemented by running a set of Perl programs for creation and usage in a semi-automatic manner. The update of the database can be done by using a specific module being launched manually that is able to import newly added sequences since the last update from UniProtKB/Swiss-Prot and then to extend the database with identified well-characterized, domain and novel tandem repeats of the proteins.

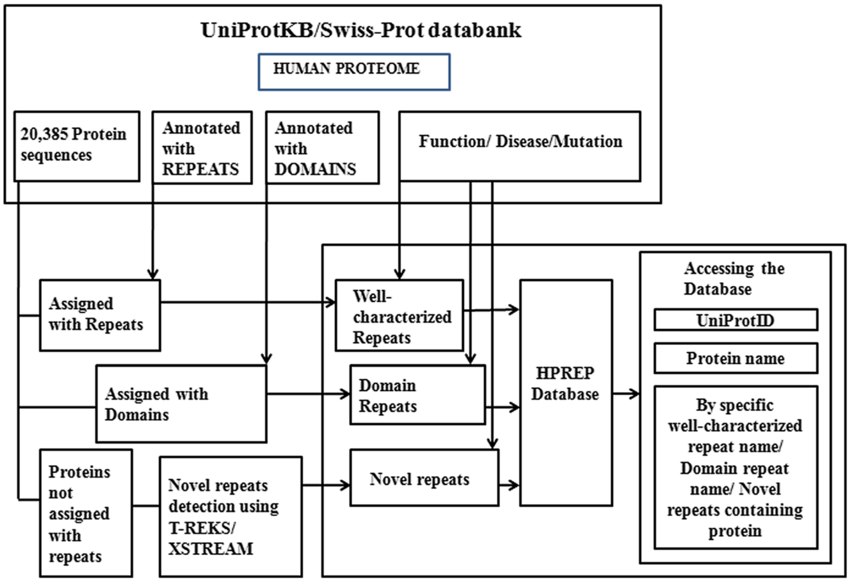

UniProtKB/Swiss-Prot databank
HUMAN PROTEOME
20,385 Protein sequences
Annotated with REPEATS
Annotated with DOMAINS
Function/ Disease/Mutation
Assigned with Repeats
Assigned with Domains
Proteins not assigned with repeats
Novel repeats detection using T-REKS/ XSTREAM
Well-characterized Repeats
Domain Repeats
Novel repeats
HPREP Database
Accessing the Database
UniProtID
Protein name
By specific well-characterized repeat name/ Domain repeat name/ Novel repeats containing protein

Figure 1: Flowchart shows the generation and accessing of HPREP database. The human protein sequences annotated with well-characterized repeats and domains repeats are collected from UniProtKB/Swiss-Prot repository and the novel repeats are identified using repeat detection tools. HPREP integrates the identified well-characterized, domain and novel repeats of human proteins with the details of functional, mutational and diseases for accessing.

2.1 Dataset collection

The complete set of 20,385 of human proteins from UniProtKB/Swiss-Prot as on 30/11/2019 with Uniprot ID, amino acid sequence, sequence length, protein name, protein family and gene as well as annotated with function, family & domain, and pathology & biotech was obtained [24] and stored in a file. The functional annotation contains the general function, active site, binding site, motif, calcium binding and nucleotide binding of the protein. The family & domain contains the details of repeat, motif, domain and zinc finger type of the protein. The repeat section includes repeated regions with specific name such as LRR, Ankyrin, Kelch and HEAT, etc., if exists otherwise with no name. The motif section contains the motif regions and their functions. The domain includes the functional domains as well as their regions. Specific annotation is available for zinc finger type protein which contains zinc finger domains and their regions. The pathology & biotech provides the mutational residues and diseases of the protein.

2.2 Identification of different categories of repeats in human proteins

2.2.1 Well-characterised repeats: The 1955 Proteins that are annotated with repeats were identified by using in-house developed Perl program. Then, the protein name, protein family, sequence, length, function, mutation and disease of the repeat proteins were retrieved from the stored information. Then, these proteins are grouped according to specific repeat names for further analysis.

2.2.2 Domain repeats: The 8446 human proteins that have functional domain assignments were analyzed for the presence of domains repeated with same family by using in-house developed Perl program. Further, the 1786 zinc-finger type proteins were also analyzed for repeats in zinc finger domains. Then, the protein name, family, sequence, length, function, mutation and disease of 3018 domains repeats identified proteins were retrieved from the stored information. Further, these proteins are grouped according to domain names for further analysis.

2.2.3 Novel tandem repeats: The UniProtKB/Swiss-Prot proteins are annotated with repeats if they have a repeated sequence motifs or repeated domains of a specific protein or protein family. However, the repeat patterns of the proteins that are not studied are uncharacterised repeats which can be identified without prior knowledge by using novel repeat detection algorithms. In this study, we used the T-REKS tool that uses *K*-means clustering algorithm as well as XSTREAM web server that uses the short seed extension method since both the methods are observed more efficient in producing true repeats from large sequence databases [11]. The novel tandem repeats of the proteins are identified by running T-REKS (http://bioinfo.montp.cnrs.fr/?r=t-reks) with the parameter settings of 20% length variability, similarity threshold of 0.7 and disallow overlaps. The XSTREAM web server (https://amnewmanlab.stanford.edu/xstream) is executed with the parameters of moderate degeneracy, high significance, a minimum word match similarity of 0.7 and redundancy removal. We set the similarity threshold to 0.7 and a minimal total length of tandem repeat of 14 residues to these two programs since the repeats based on this approximation is considered as true repeats with potential biological meaning [11].

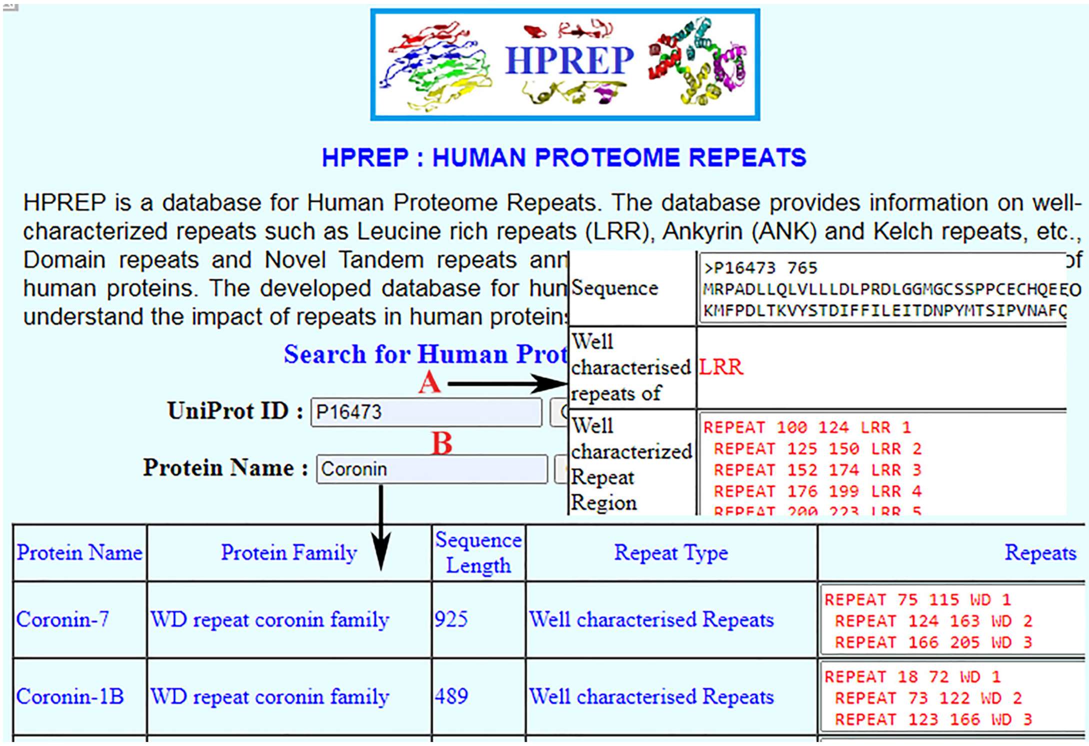
2.3 Development of database for human proteome repeats

The obtained repeats are categorized into (i) well-characterized repeats of 84 different repeat types and 320 other repeats with no specific name, (ii) domain repeats of 193 functional domains and 37 zinc finger domains and (iii) novel tandem repeats and available as a web database HPREP developed by using Perl and HTML. The developed database displays the different categories of human repeats along with functional, mutational and disease information of the proteins for further analysis. Figure 1 shows the generation and accession of the database in the form of a flow chart.

3 Results

3.1 Description of the database

The HPREP database can be accessed via the web link http://bioinfo.bdu.ac.in/hprep. Figure 2 shows the home page that includes the details about the database with search option for repeats either by using UniPort ID or by using protein name. Figure 2A shows the LRR repeats (100–124/125–150/152–174/176–199/200–223/227–248/250–271) and their sequence region, function, motif, mutation and disease of the Thyrotropin receptor protein obtained by giving UniProt ID (P16473) of the protein. The search of the database by using the protein name (Coronin) (Figure 2B) displays the list of seven Coronin names containing proteins with WD repeats in which



HPREP

HPREP : HUMAN PROTEOME REPEATS

HPREP is a database for Human Proteome Repeats. The database provides information on well-characterized repeats such as Leucine rich repeats (LRR), Ankyrin (ANK) and Kelch repeats, etc., Domain repeats and Novel Tandem repeats ann of human proteins. The developed database for hun understand the impact of repeats in human protein

Search for Human Prot

UniProt ID : P16473

Protein Name : Coronin

Sequence	>P16473 765 MRPADLLQLVLLLDLPRDLGGMGCSSPPCECHQEEO KMFPDLTKVYSTDIFFILEITDNPYMTSIPVNAFQ
Well characterised repeats of	LRR
Well characterized Repeat Region	REPEAT 100 124 LRR 1 REPEAT 125 150 LRR 2 REPEAT 152 174 LRR 3 REPEAT 176 199 LRR 4 REPEAT 200 223 LRR 5

Protein Name	Protein Family	Sequence Length	Repeat Type	Repeats
Coronin-7	WD repeat coronin family	925	Well characterised Repeats	REPEAT 75 115 WD 1 REPEAT 124 163 WD 2 REPEAT 166 205 WD 3
Coronin-1B	WD repeat coronin family	489	Well characterised Repeats	REPEAT 18 72 WD 1 REPEAT 73 122 WD 2 REPEAT 123 166 WD 3

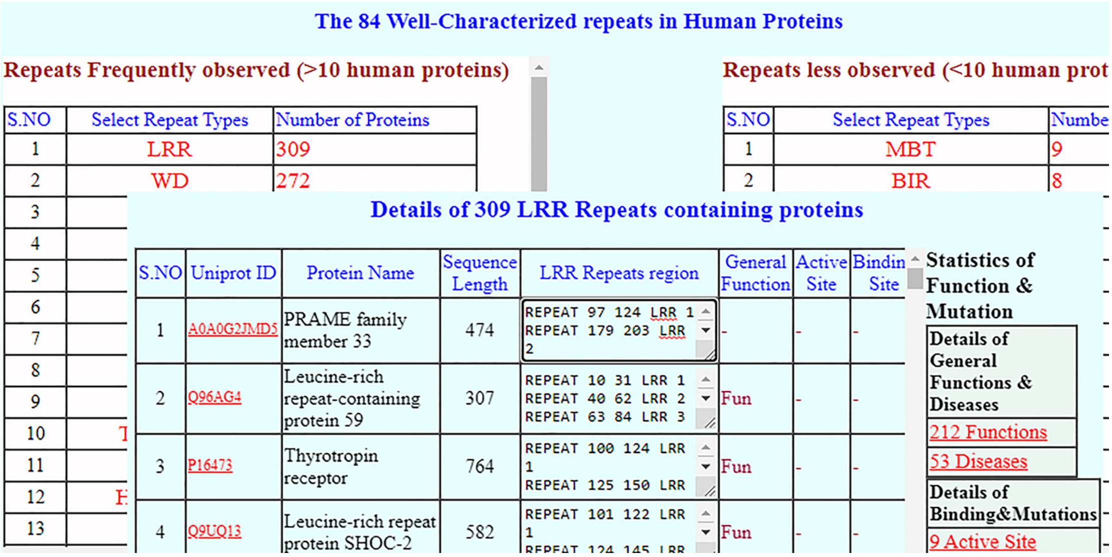
Figure 2: The homepage of HPREP database is shown with (A) the repeats of Thyrotropin receptor protein that was obtained by the search using UniProt ID (P16473) of the protein and (B) the list of seven Coronin name containing proteins with repeat details that was obtained by a search using protein name (Coronin).

further details can be obtained by clicking corresponding UniProt ID link. The database also contains the different categories of repeats of (i) well-characterized repeats (ii) domain repeats and (iii) novel tandem repeats of human proteins to search by their specific repeat name. Further, the analysis of numbers of human proteins with repeats shows an approximately of 25% of them contained repeats.

3.2 Preliminary analysis of different categories of human proteins in the database

3.2.1 Search by well-characterised repeats

The observed 84 different repeat types and their number of occurrences has suggested the abundant occurrences of LRR repeats (309 proteins), WD repeats (272 proteins), ANK repeats (250 proteins) and TPR (155 proteins) repeats in human proteins compared to other types (Figure 3). The details of specific repeat containing protein's functional and diseases can be obtained by selecting the desired one (For e.g. LRR repeats). For example, the 309 LRR repeats proteins' UniProt ID, protein name, sequence length, LRR repeats region as well as whether these proteins with general functions, active sites, binding sites, motifs, calcium binding, nucleotide binding, mutations and diseases are shown in Figure 3. From Figure 3, we observed that Nucleotide-Binding Domain, Leucine-Rich Repeat Proteins (NLP), Preferentially Expressed Antigen in Melanoma (PRAME), and Toll-like receptor family proteins containing these repeats as single pair to multiple copies of 30 with 20–30 residues in length. The number of LRR proteins with functions (202 proteins), diseases (53 proteins), active sites (9 proteins), binding sites (8 proteins), motifs (17 proteins), calcium bindings (1 protein), nucleotide bindings (25 proteins) and mutations (62 proteins) are displayed with links on the right side of the page to view their details. The 212 functional details show the diverse functions of LRR proteins of



The 84 Well-Characterized repeats in Human Proteins

Repeats Frequently observed (>10 human proteins)

S.NO	Select Repeat Types	Number of Proteins
1	LRR	309
2	WD	272
3		
4		
5		
6		
7		
8		
9		
10		
11		
12		
13		

Repeats less observed (<10 human prot

S.NO	Select Repeat Types	Numbe
1	MBT	9
2	BIR	8

Details of 309 LRR Repeats containing proteins

S.NO	Uniprot ID	Protein Name	Sequence Length	LRR Repeats region	General Function	Active Site	Bindin Site
1	A0A0G2JMD5	PRAME family member 33	474	REPEAT 97 124 LRR 1 REPEAT 179 203 LRR 2	-	-	-
2	Q96AG4	Leucine-rich repeat-containing protein 59	307	REPEAT 10 31 LRR 1 REPEAT 40 62 LRR 2 REPEAT 63 84 LRR 3	Fun	-	-
3	P16473	Thyrotropin receptor	764	REPEAT 100 124 LRR 1 REPEAT 125 150 LRR	Fun	-	-
4	Q9UQ13	Leucine-rich repeat protein SHOC-2	582	REPEAT 101 122 LRR 1 REPEAT 124 145 LRR	Fun	-	-

Statistics of Function & Mutation

Details of General Functions & Diseases
212 Functions
53 Diseases

Details of Binding&Mutations
9 Active Site

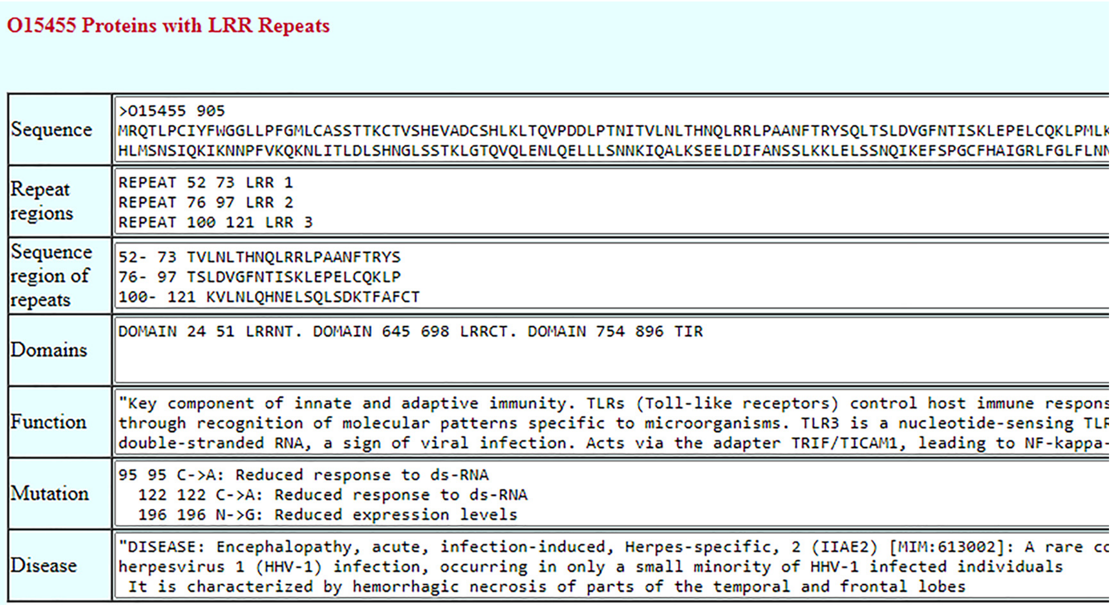
Figure 3: The different well-characterized repeats of human proteins and the details of LRR repeat containing proteins functions, mutations and diseases.

signal transduction, cell adhesion and DNA damage repair. While analysis the functional regions of active site, binding site, motif, calcium binding and nucleotide binding in LRR, we observed binding site in the LRR region such as the Toll-like receptor nine protein (Q9NR96) performs cytidine-phosphateguanosine (CpG)-DNA binding through 132 and 208 residues which are in the LRR region of (122–147 and 198–221). However, most of them are not in the LRR region.

Further, the associated diseases of LRR proteins such as myopia, epilepsy, rheumatoid arthritis and sclerosis are found out. We observed 17 mutations out of 62 mutations in the repeat regions such as the mutations (358 C->A: 358 C->S) for loss of binding with CRY1 (Cryptochrome Circadian Regulator 1) in LRR repeats (119–146/181–207/208–233/234–259/316–341/343–368/369–394) of F-box/LRR-repeat protein 3 (Q9UKT7). Further, the involvement of repeats in diseases of proteins is also observed such as Toll-like receptor 3 (O15455) protein with mutations in the regions (95 C->A: 122 C->A: 196 N->G: 247 N->R) of LRR repeats cause reduced response to double-stranded RNA that lead to acute infection-induced (herpes-specific) encephalopathy-2 (IIAE2) (Figure 4). This suggests that LRR repeats are for general protein function and sometimes get mutated for causing diseases. Likewise, the roles of well-characterized repeats can be observed by using the developed database.

3.2.2 Search by using a specific domain name

The obtained functional domain repeats and their number of occurrences shows the abundant occurrence of Ig-like C2-type followed by EF-hand, Fibronectin type-III and Cadherin in human proteins. Further, the details of specific domain proteins can be obtained by selecting the desired one. The details of 205 Ig-like C2-type proteins have shown that Carcinoembryonic antigen-related cell adhesion molecule, Leukocyte immunoglobulin-like receptor subfamily A member 3 and Vascular cell adhesion proteins are generally containing these repeats as single pair to multiple copies of 42 with 72–100 residues in length. Further, the number of Ig-like C2-type proteins containing functions and diseases can be obtained. The 172 functional details show the cell–cell adhesion and immune response modulation functions of the proteins. From the analysis of functional regions in these repeats, we observed motifs in Ig-like C2-type region such as EWI motif (250–252) of Immunoglobulin superfamily member 3 (O75054) in Ig-like C2-type region of 143–262. Further, the



O15455 Proteins with LRR Repeats

Sequence	>O15455 905 MRQTLPCIYFWGGLLPFGMLCASSTTKCTVSHEVADCSHLKLTQVPDDLPTNITVLNLTHNQLRRLPAANFTRYSQLTSLDVGFNTISKLEPELCQKLPML HLMSNSIQKIKNNPFVKQKNLITLDLSHNGLSSTKLGTQVQLENLQELLLSNNKIQALKSEELDIFANSSLKKLELSSNQIKEFSPGCFHAIGRLFGLFLN
Repeat regions	REPEAT 52 73 LRR 1 REPEAT 76 97 LRR 2 REPEAT 100 121 LRR 3
Sequence region of repeats	52- 73 TVLNLTHNQLRRLPAANFTRYS 76- 97 TSLDVGFNTISKLEPELCQKLP 100- 121 KVLNLQHNELSQLSDKTFAFCT
Domains	DOMAIN 24 51 LRRNT. DOMAIN 645 698 LRRCT. DOMAIN 754 896 TIR
Function	"Key component of innate and adaptive immunity. TLRs (Toll-like receptors) control host immune respons through recognition of molecular patterns specific to microorganisms. TLR3 is a nucleotide-sensing TL double-stranded RNA, a sign of viral infection. Acts via the adapter TRIF/TICAM1, leading to NF-kappa-
Mutation	95 95 C->A: Reduced response to ds-RNA 122 122 C->A: Reduced response to ds-RNA 196 196 N->G: Reduced expression levels
Disease	"DISEASE: Encephalopathy, acute, infection-induced, Herpes-specific, 2 (IIAE2) [MIM:613002]: A rare c herpesvirus 1 (HHV-1) infection, occurring in only a small minority of HHV-1 infected individuals It is characterized by hemorrhagic necrosis of parts of the temporal and frontal lobes

Figure 4: The Toll-like receptor 3 (O15455) protein with mutations in the residues (95 C->A: 122 C->A: 196 N->G: 247 N->R) of LRR repeats cause the reducing the response to double-stranded RNA that lead to acute infection-induced (herpes-specific) encephalopathy-2 (IIAE2).

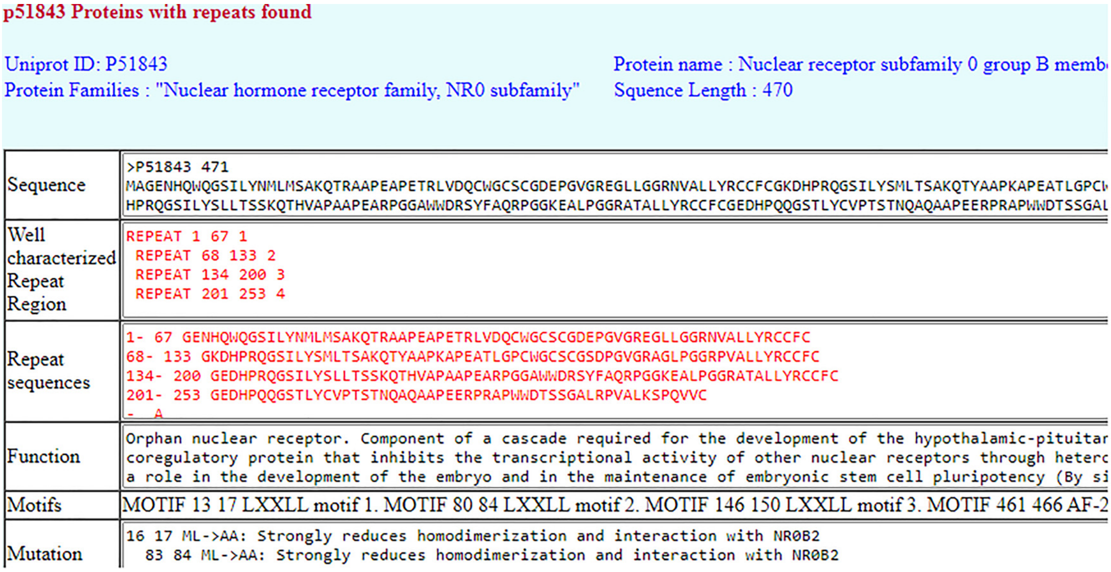
associated diseases of cardiomyopathy, hearing loss and cancer in the bladder and prostatic of the repeat proteins are observed. The details of 52 mutations show 15 mutations in the repeats such as mutations (63 E->R: 66 E->R: 84 T->R) for binding growth factor GAS6 in the repeat region (27–128/139–222) of tyrosine-protein kinase receptor (P30530) protein. Likewise, the functions and diseases of specific domain proteins as well as their involvement in functions and mutations are found out using the developed database.

3.2.3 Search for Novel repeats

The 132 tandem repeats of the proteins apart from well annotated repeats of UniProtKB using T-REKS and 202 proteins with repeats using XSTREAM can be obtained. While analysis of overlapping or completely different set, the 92 protein repeats that cover nearly the same regions in both programs are observed because of the closest definition of similarity threshold and minimal total length are given to these programs. The proteins with novel repeats along with other information of whether they have general function, active site, binding site, motif, calcium binding, nucleotide binding, mutation and diseases can be obtained by clicking the corresponding links.

3.2.4 Comparison with other database

We compared the performance of our database with PRDB database (http://bioinfo.montp.cnrs.fr/?r=repeatDB) which contains tandem repeats of UniProtKB/Swiss-Prot detected by using T-REKS program. We observed that PRDB contains 848 human tandem repeats of length >14 with entries for each repeat of protein, while in our database includes 672 proteins with repeats covering the same regions of UniProtKB annotated repeats as well as novel tandem repeats. For example, PRDB shows the repeats (8–273) with general function, Gene ontology, subcellular localization, pfam domain of the Nuclear receptor subfamily 0 group B member 1 (P51843) protein whereas, our database shows well-characterised repeats (1–253), gene, domain,



p51843 Proteins with repeats found

Uniprot ID: P51843
Protein Families : "Nuclear hormone receptor family, NR0 subfamily"

Protein name : Nuclear receptor subfamily 0 group B memb
Squence Length : 470

Sequence	>P51843 471 MAGENHQWQGSILYNMLMSAKQTRAAPEAPETRLVDQCWGCSCGDEPGVGREGLLGGRNVALLYRCCFCGKDHPRQGSILYSMLTSAKQTYAAPKAPEATLGPC HPRQGSILYSLLTSSKQTHVAPAAPEARPGGAWWDRSYFAQRPGGKEALPGGRATALLYRCCFCGEDHPQQGSTLYCVPTSTNQAQAAPEERPRAPWWDTSSGAL
Well characterized Repeat Region	REPEAT 1 67 1 REPEAT 68 133 2 REPEAT 134 200 3 REPEAT 201 253 4
Repeat sequences	1- 67 GENHQWQGSILYNMLMSAKQTRAAPEAPETRLVDQCWGCSCGDEPGVGREGLLGGRNVALLYRCCFC 68- 133 GKDHPRQGSILYSMLTSAKQTYAAPKAPEATLGPCWGCSCGSDPGVGRAGLPGGRPVALLYRCCFC 134- 200 GEDHPRQGSILYSLLTSSKQTHVAPAAPEARPGGAWWDRSYFAQRPGGKEALPGGRATALLYRCCFC 201- 253 GEDHPQQGSTLYCVPTSTNQAQAAPEERPRAPWWDTSSGALRPVALKSPQVVC - A
Function	Orphan nuclear receptor. Component of a cascade required for the development of the hypothalamic-pituitar coregulatory protein that inhibits the transcriptional activity of other nuclear receptors through heter a role in the development of the embryo and in the maintenance of embryonic stem cell pluripotency (By si
Motifs	MOTIF 13 17 LXXLL motif 1. MOTIF 80 84 LXXLL motif 2. MOTIF 146 150 LXXLL motif 3. MOTIF 461 466 AF-2
Mutation	16 17 ML->AA: Strongly reduces homodimerization and interaction with NR0B2 83 84 ML->AA: Strongly reduces homodimerization and interaction with NR0B2

Figure 5: The output repeat result obtained for a Nuclear receptor subfamily 0 group B member 1 (P51843) by using HRPEP database.

general function, disease, LXXLL motifs (13–17/80–84/146–150) for transcription factor binding and mutations (16–17 (ML->AA); 83–84 (ML->AA); 149–150 (LL->AA)) which inhibit the transcriptional activity of the protein (Figure 5). Further, the PRDB shows the consensus patterns and structure forming potential of the repeats as well as the similar search repeats in PRDB which are useful for functional analysis. Such details from PRDB could be used to enrich the HPREP database in the future for better analysis of human repeats.

4 Conclusion

The human proteome comprises repeats in nearly one third of the proteins and significant portions of repeats proteins carrying fundamental functional roles have been observed. Furthermore, the high incidence of repeats in virulence factors, amyloidogenic, prion and other disease-related sequences of the proteins has suggested that repeats are not only performing biological functions but also related to number of human diseases. The database HPREP for human proteome repeats has been developed with the aim of understanding the different categories of repeats in human proteins and their involvement in function, mutation and disease of the proteins. This knowledge could be used for better understanding the underlying roles of repeats in human proteins for drug development and identify promising biomolecules for diagnostic and prognostic purposes.

Author contribution: All the authors have accepted responsibility for the entire content of this submitted manuscript and approved submission.

Research funding: Dr D. Mary Rajathei acknowledges University Grants Commission, India for a UGC Post-Doctoral Fellow for Women Research Grant No: F.15-1/2017-18/PDFWM-2017-18-TAM-44286/(SA-II).

Conflict of interest statement: The authors declare no conflicts of interest regarding this article.

References

1. Luo H, Nijveen H. Understanding and identifying amino acid repeats. Brief Bioinform 2014;15:582–91.
2. Matsushima N, Enkhbayar P, Kamiya M, Osaki M, Kretsinger RH. Leucine–Rich Repeats (LRRs): structure, function, evolution and interaction with ligands. Drug Des Rev 2005;2:305–22.
3. Batrukova MA, Betin VL, Rubtsov AM, Lopina OD. Ankyrin: structure, properties, and functions. Biochemistry 2000;65: 395–408.
4. Coates JC. Armadillo repeat proteins: beyond the animal kingdom. Trends Cell Biol 2003;13:463–71.
5. Vetting MW, Hegde SS, Fajardo JE, Fiser A, Roderick SL, Takiff HE, et al. Pentapeptide repeat proteins. Biochemistry 2006;45: 1–10.
6. Sawaya MR, Wojtowicz WM, Andre I, Qian B, Wu W, Baker D, et al. A double shape provides the structural basis for the extraordinary binding specificity of Dscam isoforms. Cell 2008;134:1007–18.
7. Elkins PA, Ho YS, Smith WW, Janson CA, D'Alessio KJ, McQueney MS, et al. Structure of the C-terminally truncated human ProMMP9, a gelatin-binding matrix metalloproteinase. Acta Crystallogr D Biol Crystallogr 2002;58:1182–92.
8. Lee MS, Gippert GP, Soman KV, Case DA, Wright PE. Three-dimensional solution structure of a single zinc finger DNA-binding domain. Science 1989;245:635–7.
9. Heger A, Holm L. Rapid automatic detection and alignment of repeats in protein sequences. Proteins 2000;41:224–37.
10. Szklarczyk R, Heringa J. Tracking repeats using significance and transitivity. Bioinformatics 2004;20:i311–17.
11. Jorda J, Kajava AV. T-REKS: identification of tandem repeats in sequences with a K-means based algorithm. Bioinformatics 2009;25:2632–8.
12. Newman A, Cooper J. Xstream: a practical algorithm for identification and architecture modeling of tandem repeats in protein sequences. BMC Bioinf 2007;8:382.
13. Pellegrini M, Renda ME, Vecchio A. Ab initio detection of fuzzy amino acid tandem repeats in protein sequences. BMC Bioinf 2012;21:13.
14. Mary Rajathei D, Selvaraj S. Analysis of sequence repeats of proteins in the PDB. Comput Biol Chem 2013;47:156–66.
15. Rajathei DM, Parthasarathy S, Selvaraj S. Identification and analysis of long repeats of proteins at the domain level. Front Bioeng Biotechnol 2019;7:250.
16. Depledge DP, Lower RP, Smith DF. Repseq – a database of amino acid repeats present in lower eukaryotic pathogens. BMC Bioinf 2007;8:122.
17. Luo H, Lin K, David A, Nijveen H, Leunissen JAM. Prorepeat: an integrated repository for studying amino acid tandem repeats in proteins. Nucleic Acids Res 2012;40:D394–9.
18. Jorda J, Baudrand T, Kajava AV. PRDB: protein repeat database. Proteomics 2012;12:1333–6.
19. Selvaraj S, Rajathei M. A web database IR-PDB for sequence repeats of proteins in the Protein Data Bank. Int J Knowl Discov Bioinf 2017;7:1–10.
20. Pellegrini M. Tandem repeats in proteins: prediction algorithms and biological role. Front Bioeng Biotechnol 2015;3:143.
21. Kajava AV, Squire JM, Parry DA. Beta-structures in fibrous proteins. Adv Protein Chem 2006;73:1–15.
22. Baxa U, Cassese T, Kajava AV, Steven AC. Structure, function, and amyloidogenesis of fungal prions: filament polymorphism and prion variants. Adv Protein Chem 2006;73:125–80.
23. Nelson R, Eisenberg D. Structural models of amyloid-like fibrils. Adv Protein Chem 2006;73:235–62.
24. UniProt Consortium. UniProt: a worldwide hub of protein knowledge. Nucleic Acids Res 2019;47:D506–15.
25. Faux N. Single amino acid and trinucleotide repeats, function and evolution. Adv Exp Med Biol 2012;769:26–40.
26. Mularoni L, Guigó R, Albà MM. Mutation patterns of amino acid tandem repeats in the human proteome. Genome Biol 2006;7: R33.
27. Orr HT, Zoghbi HY. Trinucleotide repeat disorders. Annu Rev Neurosci 2007;30:575–621.
28. Messaed C, Rouleau GA. Molecular mechanisms underlying polyalanine diseases. Neurobiol Dis 2009;34:397–405.
29. Lieberman AP, Shakkottai VG, Albin RL. Polyglutamine repeats in neurodegenerative diseases, annual review of pathology. Mechanisms of Disease 2019;14:1–27.
30. Li C, Nagel J, Androulakis S, Lupton CJ, Song J, Buckle AM. PolyQ 2.0: an improved version of PolyQ, a database of human polyglutamine proteins. Oxford: Database; 2016;2016:1–8.
31. Karlin S, Brocchieri L, Bergman A, Mrazek J, Gentles AJ. Amino acid runs in eukaryotic proteomes and disease associations. Proc Natl Acad Sci USA 2002;99:333–8.
32. Albà MM, Guigó R. Comparative analysis of amino acid repeats in rodents and humans. Genome Res 2004;14:549–54.