BMC Bioinformatics

Database

ProtRepeatsDB: a database of amino acid repeats in genomes

Mridul K Kalita†[1], Gowthaman Ramasamy†[1], Sekhar Duraisamy†[2], Virander S Chauhan[3] and Dinesh Gupta*[1]

Address: [1]Structural and Computational Biology Group, Malaria Group, International Centre for Genetic Engineering and Biotechnology (ICGEB), Aruna Asaf Ali Marg, New Delhi 110067, India, [2]Dana-farber Cancer Institute, Harvard Medical School, Dana-830, 44-Binney street, Boston, MA-02115, USA and [3]Malaria Group, International Centre for Genetic Engineering and Biotechnology (ICGEB), Aruna Asaf Ali Marg, New Delhi 110067, India

Email: Mridul K Kalita - mridulkk@icgeb.res.in; Gowthaman Ramasamy - gowtham@icgeb.res.in; Sekhar Duraisamy - sekhar_duraisamy@dfci.harvard.edu; Virander S Chauhan - virander@icgeb.res.in; Dinesh Gupta* - dinesh@icgeb.res.in

* Corresponding author †Equal contributors

Published: 07 July 2006

BMC Bioinformatics 2006, **7**:336 doi:10.1186/1471-2105-7-336

Received: 22 February 2006
Accepted: 07 July 2006

This article is available from: http://www.biomedcentral.com/1471-2105/7/336

Abstract

Background: Genome wide and cross species comparisons of amino acid repeats is an intriguing problem in biology mainly due to the highly polymorphic nature and diverse functions of amino acid repeats. Innate protein repeats constitute vital functional and structural regions in proteins. Repeats are of great consequence in evolution of proteins, as evident from analysis of repeats in different organisms. In the post genomic era, availability of protein sequences encoded in different genomes provides a unique opportunity to perform large scale comparative studies of amino acid repeats. ProtRepeatsDB http://bioinfo.icgeb.res.in/repeats/ is a relational database of perfect and mismatch repeats, access to which is designed as a resource and collection of tools for detection and cross species comparisons of different types of amino acid repeats.

Description: ProtRepeatsDB (v1.2) consists of perfect as well as mismatch amino acid repeats in the protein sequences of 141 organisms, the genomes of which are now available. The web interface of ProtRepeatsDB consists of different tools to perform repeat s; based on protein IDs, organism name, repeat sequences, and keywords as in FASTA headers, size, frequency, gene ontology (GO) annotation IDs and regular expressions (REGEXP) describing repeats. These tools also allow formulation of a variety of simple, complex and logical queries to facilitate mining and large-scale cross-species comparisons of amino acid repeats. In addition to this, the database also contains sequence analysis tools to determine repeats in user input sequences.

Conclusion: ProtRepeatsDB is a multi-organism database of different types of amino acid repeats present in proteins. It integrates useful tools to perform genome wide queries for rapid screening and identification of amino acid repeats and facilitates comparative and evolutionary studies of the repeats. The database is useful for identification of species or organism specific repeat markers, interspecies variations and polymorphism.

Background

The function, type and size of amino acid repeat regions in proteins are vastly diverse. The amino acid repeats (hence forth referred as repeats and repeat containing proteins as repeat proteins) can be functionally neutral or active, tandem or scattered, and perfect or mismatch. Repeats may also range from a very minuscule to a highly significant fraction of proteins. Though functions of a few types of repeats are known, in general, function and evolution of a large variety of repeats are still poorly understood. Repeats have been broadly classified as perfect and mismatch repeats. The former class is further subdivided into homopeptide repeats (reiteration of any single amino acid, henceforth referred to as homo repeats) and heteropeptide repeats (repeats with different amino acids, henceforth referred to as hetero repeats). The mismatch repeats, consisting of repeats with substituted conserved amino acids, form the most diverse class of repeats.

Homo repeats containing proteins have been recognized as cause of several neurodegenerative and congenital malformation diseases. Proteins containing polyglutamine stretches (polyQ) have been observed in at least eight neural diseases [1]. Similarly, proteins with polyalanine tracts (polyA) are associated with various congenital malformations, skeletal dysplasia and nervous system anomalies [2]. Most of such homo repeats arise from trinucleotide slippage during replication [3,4], leading to proteins with homopolymeric tracts. Such proteins are either susceptible to misfolding, or aggregation and subsequent degradation, rendering them impaired.

Hetero repeats in proteins are quite diverse. For example, the glycine-proline amino acid run in eukaryotic genomes, polar zippers and prion like glutamine/asparagine-rich stretches [5-7], whereas mucins contain arrays of tandem repeats which are rich in serine and threonine residues corresponding to the most O-glycosylated part of the mature protein [8].

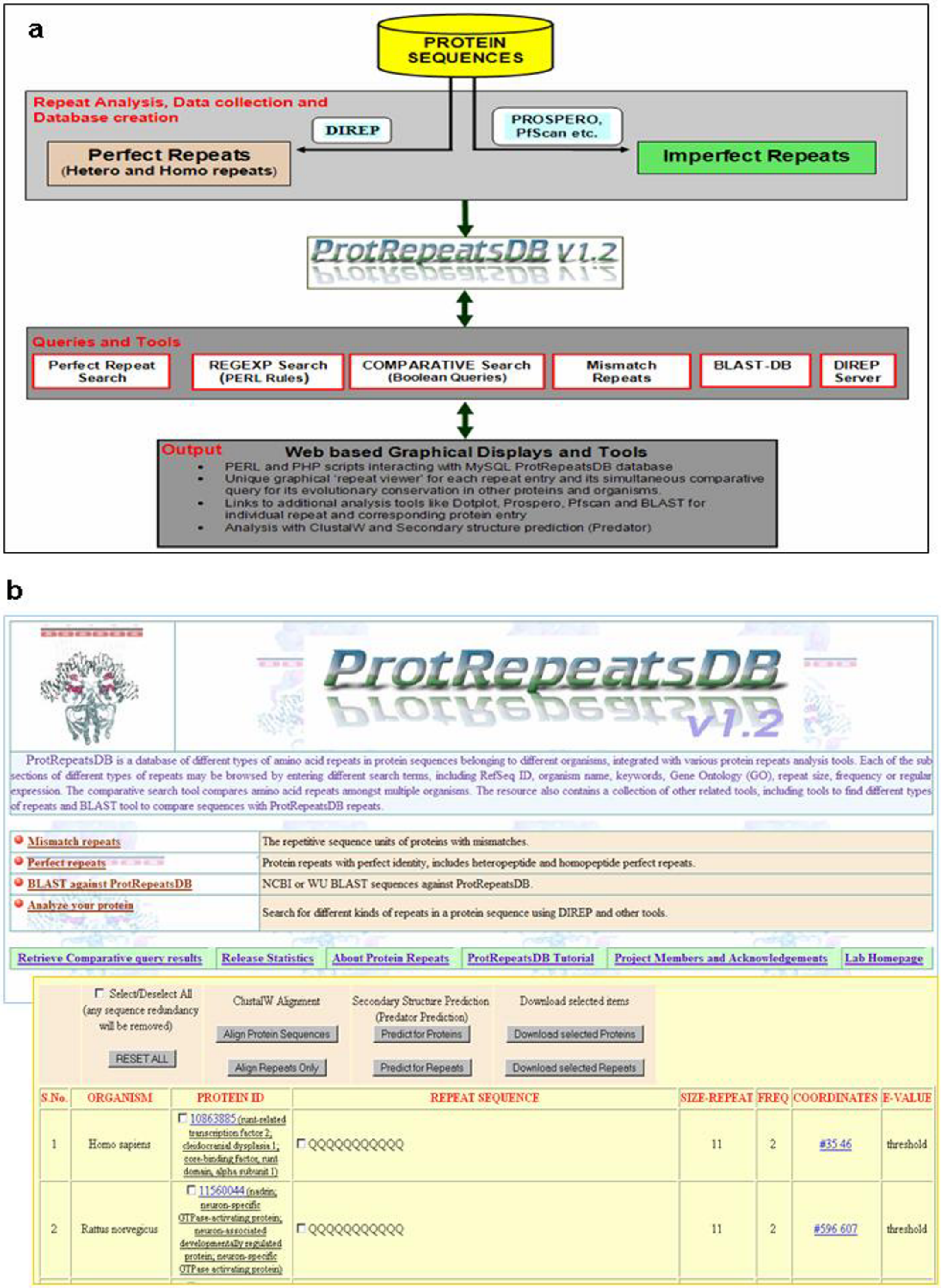
Functional studies have shown that the insertions of mismatch repetitive regions unique to a species can be indispensable to the functions of proteins, for example as shown in *Plasmodium berghei* glucose-6-phosphate dehydrogenase-6-phosphogluconolactonase, a bi-functional enzyme [9]. The role of repeats in antigenic variation and immune evasion by pathogens (e.g. *Plasmodium falciparum*, *Trypanosoma brucei* and group B *Streptococci*) has been elucidated [10-14]. Certain repeats provide structural scaffolds for structural packing of functional groups of amino acids to facilitate molecular interactions and stabilization. For example, Yeast Sup35p protein contains five oligopeptide repeats, which stabilizes its aggregation [15].

Identification of perfect repeats is simpler than that of mismatch repeats. All strategies to identify mismatch repeats have some limitation or the other, on account of polymorphism, low similarity and vast diversity of repeats [16]. Most strategies utilize algorithms based on Smith-Waterman local alignment, using substitution matrices [17,18], but some are also aimed to locate low complexity regions of proteins [19]. It is however clear that no single algorithm or strategy is sufficient to find all types of repeats in protein sequences.

Large-scale comparative studies of repeats across kingdoms will be required for better understanding of the role, diversity and evolution of repetitive sequences in proteins. Such comparisons may also provide deeper insight into the role of repeats in folding of proteins, immunogenicity and relevance to disease etiology [20]. However, most studies on repeats, so far, have focused either on a few classes of proteins in a limited number of organisms or only a particular type of repeat in different proteins or organisms [21,22]. Tools which can perform large scale cross species comparisons of different types of repeats are not yet developed. Here we present our efforts to build a robust protein repeats database and tools for large-scale analysis of repeats in different organisms. ProtRepeatsDB is a simple relational database to facilitate complex and large-scale comparative studies of repeats amongst organisms.

Construction and contents

ProtRepeatsDB is based on an underlying pipeline, shown in the Figure 1a. Repeats were identified using a script that automates sequence retrieval and execution of various repeat finders. All protein sequences used in the present study were obtained from the Reference Sequences (RefSeq) database [23] of NCBI (National Center for Biotechnology Information). It has been reported that for a protein of average size and composition, a run of an individual amino acid is statistically significant only if it is five or more residues long [24]. Hence, we used a cutoff size of five to identify repeats. A PERL program (DIREP) was used to detect tandem or scattered- hetero and homo repeats of size greater than or equal to five. Similarly, a different PERL code was used to detect homo repeats of size greater than five and less than ten. The mismatch repeats were identified using PROSPERO [25], which performs comparison of each sequence to itself, and prints all self alignments with p-values less than a predefined threshold. PROSITE patterns or profiles representing different amino acid repeats were identified using PFSCAN [26]. All the identified repeats and corresponding details were stored in a MySQL database. A large number of entries in the database were redundant, as several internal repeats which are part of larger repeats are detected as unique repeats. For example in the knob associated histidine rich protein (RefSeq:NP_472949) of *P. falciparum*, the repeats like KKKSK, KKKSKK, KKKSKKHKDN etc. were identified as separate repeats- whereas the longest repeat alone represents all the internal repeats. However, since such inter-


a
PROTEIN SEQUENCES
Repeat Analysis, Data collection and Database creation
DIREP
PROSPERO, PfScan etc.
Perfect Repeats (Hetero and Homo repeats)
Imperfect Repeats
ProtRepeatsDB v1.2
Queries and Tools
Perfect Repeat Search
REGEXP Search (PERL Rules)
COMPARATIVE Search (Boolean Queries)
Mismatch Repeats
BLAST-DB
DIREP Server
Output
Web based Graphical Displays and Tools
b
ProtRepeatsDB v1.2

- PERL and PHP scripts interacting with MySQL ProtRepeatsDB database
- Unique graphical 'repeat viewer' for each repeat entry and its simultaneous comparative query for its evolutionary conservation in other proteins and organisms.
- Links to additional analysis tools like Dotplot, Prospero, Pfscan and BLAST for individual repeat and corresponding protein entry
- Analysis with ClustalW and Secondary structure prediction (Predator)

ProtRepeatsDB is a database of different types of amino acid repeats in protein sequences belonging to different organisms, integrated with various protein repeats analysis tools. Each of the sub sections of different types of repeats may be browsed by entering different search terms, including RefSeq ID, organism name, keywords, Gene Ontology (GO), repeat size, frequency or regular expression. The comparative search tool compares amino acid repeats amongst multiple organisms. The resource also contains a collection of other related tools, including tools to find different types of repeats and BLAST tool to compare sequences with ProtRepeatsDB repeats.

● Mismatch repeats	The repetitive sequence units of proteins with mismatches.
● Perfect repeats	Protein repeats with perfect identity, includes heteropeptide and homopeptide perfect repeats.
● BLAST against ProtRepeatsDB	NCBI or WU BLAST sequences against ProtRepeatsDB.
● Analyze your protein	Search for different kinds of repeats in a protein sequence using DIREP and other tools.

Retrieve Comparative query results
Release Statistics
About Protein Repeats
ProtRepeatsDB Tutorial
Project Members and Acknowledgements
Lab Homepage
□ Select/Deselect All (any sequence redundancy will be removed)
RESET ALL
ClustalW Alignment
Align Protein Sequences
Align Repeats Only
Secondary Structure Prediction (Predator Prediction)
Predict for Proteins
Predict for Repeats
Download selected items
Download selected Proteins
Download selected Repeats

S.No.	ORGANISM	PROTEIN ID	REPEAT SEQUENCE	SIZE-REPEAT	FREQ	COORDINATES	E-VALUE
1	Homo sapiens	□ 10863885 (runt-related transcription factor 2; cleidocranial dysplasia 1; core-binding factor, runt domain, alpha subunit 1)	□ QQQQQQQQQQQ	11	2	#35 46	threshold
2	Rattus norvegicus	□ 11560044 (nadrin; neuron-specific GTPase-activating protein; neuron-associated developmentally regulated protein; neuron-specific GTPase activating protein)	□ QQQQQQQQQQQ	11	2	#596 607	threshold

Figure 1
(a) Simplified schema of ProtRepeatsDB pipeline. (b) Snapshots of ProtRepeatsDB main page display, overlapped with the output obtained for analysis of the homo repeats in androgen receptors, see text for details.

nal repeats may be conserved in other proteomes, this redundancy was removed by performing comparison of each repeat against the entire database for occurrence other than the self. The internal repeats were retained in the database only if present in at least one protein other than the self protein containing the repeat, otherwise only the longest stretch of the repeat, which represents all the internal repeats, was retained. This procedure significantly reduced the redundancy as well as the size of the database, without loosing any information regarding conservation of the repeats in different sequences and organisms.

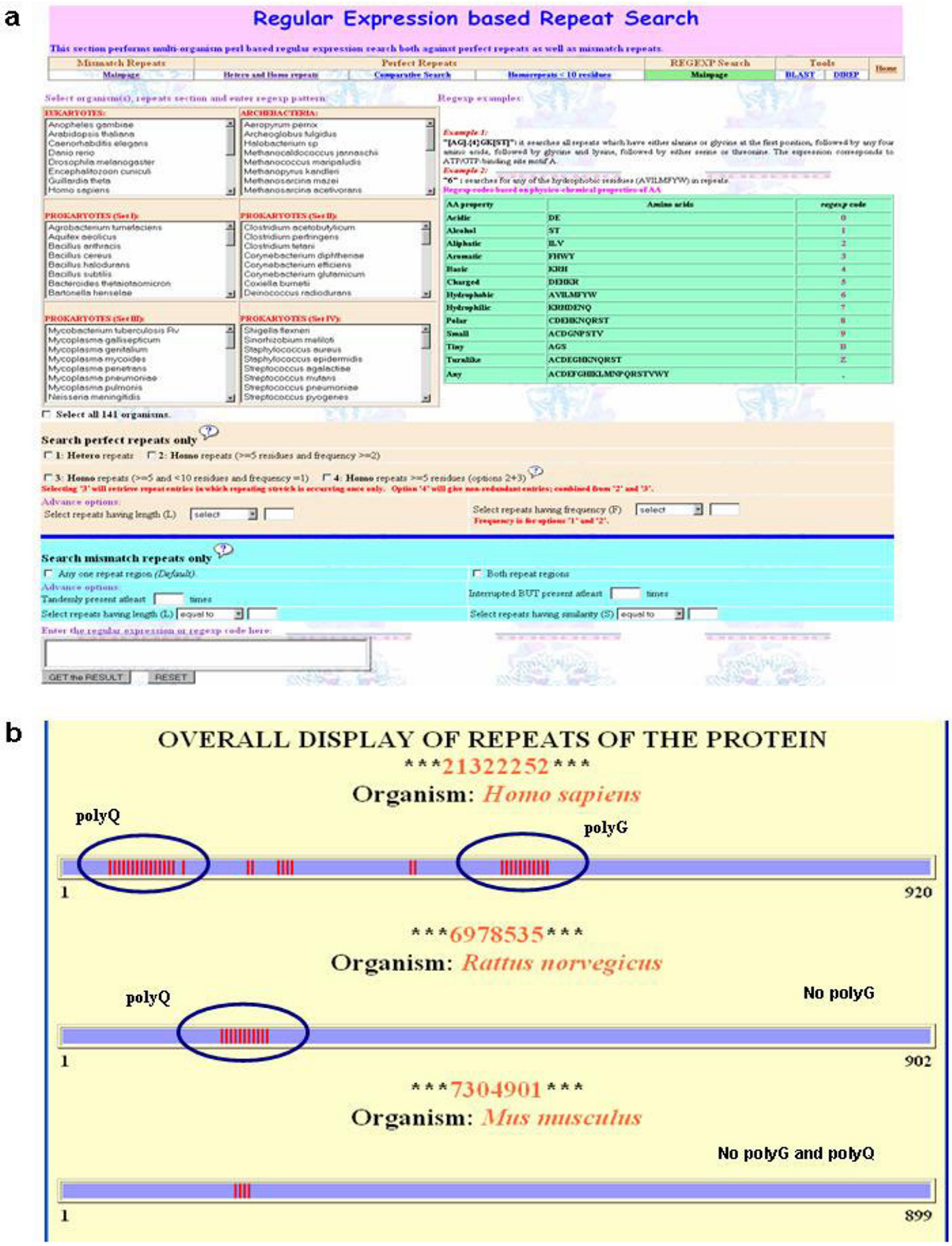
Utility and discussion

The ProtRepeatsDB and related tools facilitate retrieval and analysis of repeats in the database through different query pages and different sections. On submitting a query, a summary page of repeats satisfying the search conditions is generated. The summary page consists of a brief description of the sequences containing the repeat, such as the organism name, protein ID (hyperlinked to corresponding NCBI RefSeq entry), amino acids constituting the repeat, repeat size, repeat frequency, positions of the repeat in protein sequence (referred to as coordinates) hyperlinked to a detailed view, and E-value etc. (Figure 1b). Upon clicking coordinates of a repeat in the summary page, detailed information of the corresponding repeat is displayed as a new page consisting of two frames. The top frame consists of a dynamically generated color graph illustrating relative distribution of the repeat in the protein, followed by the sequence of the protein with repeats highlighted in red. In addition to this, links are provided to retrieve the sequence in FASTA format, perform repeat analysis using PROSPERO [25], DOTMATCHER [27], PfScan and BLAST [28] against ProtRepeatsDB. This section also provides links to display PROSITE matches, mismatch repeats and graphical representation of all the identified repeats in the protein sequence. The lower frame lists all the ProtRepeatsDB sequence entries sharing the same repeat.

The "REGEXP search" section accepts user-defined regular expressions to query the database. The regular expressions may either include patterns of single letter amino acid codes or patterns of predefined group codes based on similar physicochemical properties of amino acids. The details of predefined codes and groups are enumerated in the regular expressions query page (Figure 2a). For example, one can search the mismatch repeat section for prion repeat patterns 'P [HQ] [GS]G{1,3}WGQ', the octapeptide tandem repeat pattern associated with many neurological disorders like spongiform encephalopathies and dementia [29-31]. We used "REGEXP search" tool to query homo repeats section for glutamine and glycine homo repeats (also referred as polyQ and polyG respectively) (Figure 2a) repeats in rat, mouse and human proteomes. In the summary output page (not shown here), the NCBI RefSeq entry (RefSeq:21322252) corresponds to human androgen receptors (ARs). The human AR protein is characterized by presence of polyQ and polyG repeats, which are associated with gene polymorphisms, cryptorchidism [32] and a risk factor for prostate cancer [33]. However, we found that polyQ is absent in mouse AR and polyG is absent in both mouse (RefSeq:7304901) and rat (RefSeq:6978535) orthologous ARs (Figure 2b).

The comparative search section provides tools to compare repeats in multiple organisms using Boolean (AND or NOT) operators. For instance, by using 'AND' operator, one can find repeats common in different sets of organisms. Similarly, the "NOT" operator can be used to obtain a list of repeats which are mutually exclusive in user defined sets of organisms. Using the tools to search PROSITE repeats section, we investigated bacterial proteins with tetratricopeptide repeat (TPR), a structural repeat motif present in a wide range of proteins [34,35] and believed to be involved in protein-protein interactions and assembly of multi protein complexes [36,37]. We found that the TPR containing orthologs of *Vibrio vulnificus* CMCP6 [RefSeq:27367908], *Vibrio parahaemolyticus* [RefSeq:28901310] and *V. vulnificus* YJ016 [RefSeq:37676035] contain tandem glutamine repeats downstream from the TPR motifs. However, multiple alignments of the sequences revealed that the glutamine run is smaller in *Vibrio cholerae* [RefSeq:15600941] or interspersed, as in *Vibrio augustum* [RefSeq:90577283] and *Shewanella oneidensis* MR-1 [RefSeq:24374614]. Significance of the glutamine homo repeats in *Vibrio* proteins is not clear; however such glutamine repeats are known to induce formation of quasi-aggregates in the early stage of amyloid protein fibrillization [38] and inhibition of protein degradation [39] in various proteins associated with human neurodegenerative disease.

Overall analysis of homo repeat proteins in different proteomes was performed by dividing them into two groups of varying repeat sizes and frequency of occurrence. First group consists of proteins with smaller homo repeats of size less than ten, occurring one or more times in a protein sequence. Second group consists of proteins with longer homo repeats of size ten and greater, occurring more than once in a protein sequence. Table 1 summarizes the comparative distribution of homo repeat proteins in different proteomes in the two groups mentioned above. The table lists first five and first three proteomes, if ordered according to decreasing percentages of proteins in the first and second group, respectively. The second group list was restricted to first three organisms only, mainly because of the fact that there are not many proteins which contain extremely large as well as repeating homo repeats. For homo repeats of each of the amino acids, the correspond-



This section performs multi-organism perl based regular expression search both against perfect repeats as well as mismatch repeats.

Select organism(s), repeats section and enter regexp pattern:

EUKARYOTES:

Anopheles gambiae
Arabidopsis thaliana
Caenorhabditis elegans
Danio rerio
Drosophila melanogaster
Encephalitozoon cuniculi
Guillardia theta
Homo sapiens

ARCHEBACTERIA:

Aeropyrum pernix
Archeoglobus fulgidus
Halobacterium sp
Methanocaldococcus jannaschi
Methanococcus maripaludis
Methanopyrus kandleri
Methanosarcina mazei
Methanosarcina acetivorans

PROKARYOTES (Set I):

Agrobacterium tumefaciens
Aquifex aeolicus
Bacillus anthracis
Bacillus cereus
Bacillus halodurans
Bacillus subtilis
Bacteroides thetaiotaomicron
Bartonella henselae

PROKARYOTES (Set II):

Clostridium acetobutylicum
Clostridium perfringens
Clostridium tetani
Corynebacterium diphtheriae
Corynebacterium efficiens
Corynebacterium glutamicum
Coxiella burnetii
Deinococcus radiodurans

PROKARYOTES (Set III):

Mycobacterium tuberculosis Rv
Mycoplasma gallisepticum
Mycoplasma genitalium
Mycoplasma mycoides
Mycoplasma penetrans
Mycoplasma pneumoniae
Mycoplasma pulmonis
Neisseria meningitidis

PROKARYOTES (Set IV):

Shigella flexneri
Sinorhizobium meliloti
Staphylococcus aureus
Staphylococcus epidermidis
Streptococcus agalactiae
Streptococcus mutans
Streptococcus pneumoniae
Streptococcus pyogenes

□ Select all 141 organisms.

Regexp examples:

Example 1:
"[AG].{4}GK[ST]": it searches all repeats which have either alanine or glycine at the first position, followed by any four amino acids, followed by glycine and lysine, followed by either serine or threonine. The expression corresponds to ATP/GTP-binding site motif A.
Example 2:
"6" : searches for any of the hydrophobic residues (AVILMFYW) in repeats.

Regexp codes based on physico-chemical properties of AA

AA property	Amino acids	regexp code
Acidic	DE	0
Alcohol	ST	1
Aliphatic	ILV	2
Aromatic	FHWY	3
Basic	KRH	4
Charged	DEHKR	5
Hydrophobic	AVILMFYW	6
Hydrophilic	KRHDENQ	7
Polar	CDEHKNQRST	8
Small	ACDGNPSTV	9
Tiny	AGS	B
Turnlike	ACDEGHKNQRST	Z
Any	ACDEFGHIKLMNPQRSTVWY	.

Search perfect repeats only

□ 1: **Hetero** repeats □ 2: **Homo** repeats (>=5 residues and frequency >=2)

□ 3: **Homo** repeats (>=5 and <10 residues and frequency =1) □ 4: **Homo** repeats >=5 residues (options 2+3)

Selecting '3' will retrieve repeat entries in which repeating stretch is occurring once only. Option '4' will give non-redundant entries; combined from '2' and '3'.

Advance options:

Select repeats having length (L) select

Select repeats having frequency (F) select

Frequency is for options '1' and '2'.

Search mismatch repeats only

□ Any one repeat region *(Default)*. □ Both repeat regions

Advance options:

Tandemly present atleast times

Interrupted BUT present atleast times

Select repeats having length (L) equal to

Select repeats having similarity (S) equal to

Enter the regular expression or regexp code here:

Figure 2

ProtRepeatsDB: (a) A snapshot of the REGEXP search tool: regular expression search can be performed on different ProtRepeatsDB sections. (b) Consolidated graphical view for all the proteins with conserved polyQ in androgen receptors (see text for more details).

Table 1: Relative distribution of homo repeats (tandem or scattered) of size <10 (occurring one or more times) and ≥10 (occurring more than once). The figures within round brackets represent the number of homo repeat proteins of corresponding amino acid and organism. The figures within square brackets represent percentage of homo repeat proteins of corresponding amino acid and organism, to that of all the homo repeat proteins in the organism.

Amino Acid		Proteins with homo repeats of size <10					Proteins with homo repeats of size ≥10		
		1st	2nd	3rd	4th	5th	1st	2nd	3rd
Acidic (Polar)	**D**	Sce(94) [13.33]	Pfa(270) [11.56]	Ncr(280) [9.18]	Dre(81) [8.7]	Ara(395) [7.7]	Sce(3) [0.43]	Pfa(6) [0.44]	Ara(3) [0.39]
	E	Hsa(996) [21.95]	Mus(761) [20.35]	Dre(187) [20.08]	Rat(654) [19.53]	Spo(53) [15.54]	Mus(28) [4.11]	Hsa(25) [2.43]	Rat(9) [1.48]
Basic (Polar)	**R**	Osa(545) [12.5]	Sav(48) [11.11]	Sco(44) [10.06]	Rat(161) [4.81]	Mus(151) [4.03]	Mus(1) [0.15]	--	--
	K	Pfa(1100) [47.11]	Lin(12) [19.6]	Spo(37) [10.8]	Dre(92) [9.88]	Sce(59) [8.37]	Ara(1) [1.43]	Ncr(3) [0.56]	Pfa(2) [0.14]
	H	Ano(153) [6.28]	Dme(282) [6.06]	Ncr(75) [2.87]	Cel(68) [2.69]	Ara(135) [2.63]	--	--	--
Polar	**S**	Lpl(25) [36.23]	Ara(1742) [33.99]	Spo(112) [32.84]	Sce(194) [27.52]	Dre(221) 23.73]	Hsa(20) [1.94]	Mus(9) [1.32]	Dme(14) [0.76]
	T	Lpl(14) [20.28]	Cel(307) [12.15]	Ncr(316) [12.09]	Dme(516) [11.11]	Ano(235) [9.65]	Dme(6) [0.33]	Ano(2) [0.35]	Rat(1) [0.16]
	N	Pfa(1675) [71.73]	Sce(88) [12.48]	Dme(465) [9.99]	Ncr(146) [5.58]	Ara(234) [4.56]	Pfa(280) [20.29]	Ncr(8) [1.49]	Dme(2) [0.11]
	Q	Dme(1582) [34.01]	Lma(25) [31.6]	Ano(576) [23.65]	Ncr(525) [20.09]	Sce(129) [18.30]	Dme(131) [7.12]	Ncr(34) [6.34]	Hsa(43) [4.18]
	C	Mus(25) [0.66]	Dre(6) [0.64]	Rat(21) [0.62]	Ano(11) [0.45]	Hsa(21) [0.44]	--	--	--
	Y	Pfa(30) [1.28]	Ype(1) [1.23]	Cel(7) [0.27]	Ano(3) [0.12]	Osa(5) [0.11]	--	--	--
Non-polar	**G**	Osa(1294) [26.69]	Ano(601) [24.68]	Dme(957) [20.57]	Ncr(454) [17.38]	Mbo(47) [15.87]	Ara(11) [1.43]	Has(10) [0.97]	Dme(8) [0.44]
	A	Mbo(191) [64.52]	Sav(222) [51.38]	Sco(223) [51.02]	Osa(1520) [34.87]	Dme(1209) [25.98]	Dme(34) [1.85]	Mus(10) [1.47]	Hsa(11) [1.07]
	V	Osa(79) [1.89]	Ano(29) [1.19]	Rat(35) [1.04]	Ara(53) [1.03]	Cel(23) [0.91]	Rat(1) [0.10]	--	--
	L	Tth(48) [49.48]	Dra(49) [42.61]	Pae(49) [32.23]	Rat(662) [19.77]	Hsa(860) [18.26]	Xfa(1) [9.09]	Ano(1) [0.18]	--
	I	Sto(7) [12.72]	Pfa(18) [0.77]	Dre(7) [0.75]	Mus(26) [0.69]	Cel(14) [0.55]	--	--	--
	P	Hsa(851) [18.07]	Osa(712) [16.33]	Mus(608) [16.21]	Cel(394) [15.59]	Rat(506) [15.11]	Ara(10) [1.30]	Has(13) [1.26]	Mus(4) [0.59]
	M	Ara(16) [0.33]	Dre(3) [0.32]	Pfa(5) [0.21]	Ano(5) [0.20]	Ncr(3) [0.11]	Ara(1) [0.13]	--	--
	F	Pfa(65) [2.78]	Ara(43) [0.83]	Cel(19) [0.75]	Mus(23) [0.61]	Osa(16) [0.37]	Mus(1) [0.15]	Osa(1) [0.13]	--
	W	Ano(3) [0.12]	Osa(3) [0.06]	Ncr(1) [0.03]	--	--	--	--	--

Note: **Ano**: *Anophelis gambiae*; **Ara**: *Arabidopsis thaliana*; **Cel**: *Caenorhabditis elegans*; **Dre**: *Danio rerio*; **Dra**: *Deinococcus radiodurans*; **Dme**: *Drosophila melanogastor*; **Hsa**: *Homo sapiens*; **Lpl**: *Lactobacillus plantarum*; **Lma**: *Leishmania major*; **Lin**: *Leptospira interrogans*; **Mus**: *Mus musculus*; **Mbo**: *Mycobacterium bovis*; **Ncr**: *Neurospora crassa*; **Osa**: *Oryza sativa*; **Pfa**: *Plasmodium falciparum*; **Pae**: *Pseudomonas aeruginosa*, **Rat**: *Rattus norvegicus*; **Sce**: *Saccharomyces cerevisiae*; **Spo**: *Schizosaccharomyces pombe*; **Sav**: *Streptomyces avermitilis*; **Sco**: *Streptomyces coelicolor*; **Sto**: *Sulfolobus tokodaii*; **Tth**: *Thermus thermophilus*; **Ype**: *Yersinia pestis*; **Xfa**: *Xylella fastidiosa* ing number of protein sequences (brackets) as percentage of all the homo repeats in the organism is listed (square brackets). Table 1 brings out certain interesting facts, for example amino acid preference for homo repeats in different organisms. The asparagine homo repeats are most abundant in *P. falciparum* proteins. In *P. falciparum* proteome, we found that 1675 proteins contain asparagine homo repeats of size less than ten, and as many as 280 proteins *P. falciparum* proteins have asparagine repeats of size ten and above, occurring more than once. The proteins with asparagine repeats constitute around 31 percent of the total annotated proteins in the parasite. The abundance of asparagines repeats has been correlated with many prion-like domains in the parasite genome [40]. We also found that the proteins with lysine homo repeats also constitute a large percentage (47 percent) of homo repeat containing proteins in *P. falciparum*, which is 21 percent of all the annotated proteins in the genome. However, there are only two *P. falciparum* proteins with lysine homo repeats in the second group. The ProtRepeatsDB comparative tools also reveal that the number of glutamic acid homo repeat proteins is higher than those containing aspartic acid in all the eukaryotes, except *P. falciparum*.

Glutamine homo repeat proteins are most abundant in *D. melanogaster*, accounted by 1582 proteins, which is 34 percent of all the homo repeat proteins in the genome. The polar glutamine repeats are primarily present in proteins that are involved in transcription-translational activities especially proteins interacting with DNA and other proteins. Proline homo repeats of size ten and greater are present in thirteen proteins of *H. sapiens*, ten *A. thaliana* proteins, and four *M. musculus* proteins. One class of such proteins includes formin-like proteins, common to the above three organisms, which are involved in processes such as morphogenesis, embryonic differentiation, cell polarity, and cytokinesis [41]. The proline-rich regions of formin-like proteins are believed to be involved in protein-protein interactions as exemplified by the crystal structure of the binding domain in formin binding protein (FBP11) with specific binding to PPLPp motifs in the formins [42].

The use of ProtRepeatsDB in comparative analysis of homo repeat proteins in related species is perhaps best illustrated by comparison of such repeats in *H. sapiens, R. norvegicus* and *M. musculus*. The most abundant homo repeat proteins in human, mouse and rat have repeats of glutamic acid, proline, leucine, alanine, glutamine, glycine, serine or lysine residues. Homo repeat proteins containing repeats of isoleucine, asparagine and valine are miniscule in *H. sapiens* and related species. Number of phenylalanine homo repeat proteins in *M. musculus* is relatively higher than those in rat and human proteomes. The percentage of cysteine, methionine and arginine homo repeat protein is almost same in all the three organisms. There is no tyrosine homo repeat protein in rat and mouse; however, there is one tyrosine homo repeat in a *H. sapiens* helicase protein. The mixed-lineage leukemia proteins of human, which are also trithorax homologs of *Drosophila*, have homo repeats of serine, glutamic acid, proline and glycine. A human and mouse ortholog of the protein has a serine run of forty two and thirty amino acid residues respectively, however the run is absent in the corresponding rat ortholog. The non trithorax homologs in human mixed-lineage leukemia proteins also lack serine runs.

The Table 1 also reveals that there is no histidine, cysteine, tyrosine, isoleucine and tryptophan homo repeat falling under second group, encoded in any of the genomes. The rare methionine homo repeats have been observed in proteins belonging to *A. thaliana*, *A. gambiae* and *P. falciparum*. In fact, a CHP-rich zinc finger protein (CHP-rich: cysteine, histidine and proline rich) of *A. thaliana* contains methionine repeat of size 11, which also happens to be the longest homo repeat of methionine in the database. Phenylalanine and tyrosine repeats are present in several *P. falciparum* hypothetical proteins and putative proteins like succinyl-CoA synthetase alpha subunit, syntaxin, Ser/Thr protein kinase, sequestrin and acid phosphatase. Glycine homo repeat proteins are ubiquitously present in almost all super kingdoms. The mycobacterial parasite *M. bovis* has the highest number of alanine homo repeat proteins, representing 64.5 percent of its homo repeat proteins and nearly 4.8 percent of annotated proteins in the genome.

Statistical analysis of ProtRepeatsDB is in agreement with the earlier observation that the percentage of perfect repeat containing repeat proteins is higher in eukaryotes than in prokaryotes and archeals (Figure 3a). Our analysis indicates that hetero repeat proteins also follow a similar trend, although, homo repeats are a predominant feature of the eukaryotic genomes (Figure 3a). Similarly, the PROSITE profiles representing repeats are mainly found in eukaryotes. The high percentage of mismatch repeats amongst all super kingdoms indicates duplication events during evolution.

Figure 3b–d, is a graphical representation of perfect repeats in representative proteomes. Figure 3b gives the distribution of repeat proteins in the eukaryotes, which indicates that *P. falciparum* has an unusual distribution of repeats in comparison with other eukaryotes; it is the only organism with more than 50% of its proteome constituted by proteins containing different kinds of repeats. Since one of the important role of repeat proteins is believed to be in immune evasion by the parasites, we investigated the nature of repeats in immuno dominant proteins in *P.*

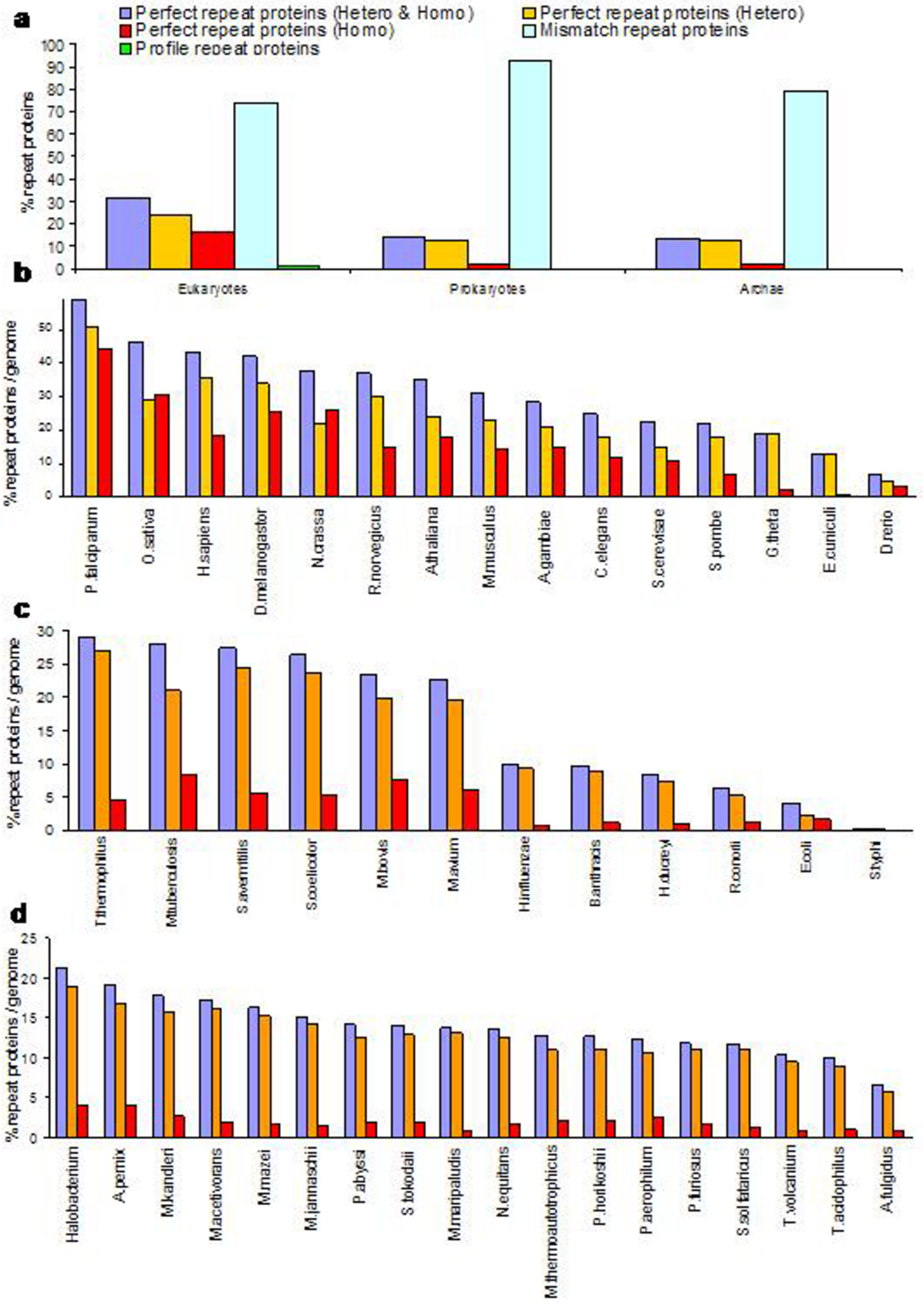


Figure 3

Comparative distribution of different types of repeat proteins in ProtRepeatsDB. (a) The percentage of repeat proteins in three super kingdoms containing perfect repeats (including hetero and homo repeats), perfect hetero repeats, perfect homo repeats, mismatch repeats and repeats with PROSITE repeat profiles. (b) The percentage of different types of perfect repeat proteins in few eukaryotic genomes in ProtRepeatsDB. (c) The percentage of different types of perfect repeat proteins in few prokaryotic genomes. (d) The percentage of different types of perfect repeat proteins in few archeal genomes. **Note**: figures b-d: The organisms with 6 highest and 6 lowest repeat protein percentages are shown here.

falciparum, and we found that the parasite antigens are not only rich in homo repeats, several of these proteins are also marked by abundance of hetero repeats. We also found that twenty one antigens of the parasite also have hetero repeats of size ten or greater. Noticeably, most of these hetero repeat regions are rich in charged or polar amino acids, which are known to play important role in antigen-antibody or antigen-carrier protein interactions. There are several such examples, including that of the S-antigen in *P. falciparum*, which is characterized by the presence of two similar repeats of sixty nine residues each, forming the coiled and surface exposed regions of the proteins. Similarly, other plasmodium antigens such as the LSA (liver stage antigen), antigen 332, MESA (mature-parasite- infected erythrocyte surface antigen) or PfEMP2 (*P. falciparum* erythrocyte membrane protein 2) and FIRA (interspersed repeat antigen) also have large inserts of hetero repeats. MESA and ring-infected erythrocyte antigen are surface exposed antigens rich in glutamic acid and lysine residues. Some of these antigens are also being developed as vaccines [43].

Amongst other eukaryotes, we found abundant repeat proteins in *O. sativa*, *H. sapiens*, *D. melanogastor* and *Neurospora crassa*. Interestingly, *N. crassa* and *O. Sativa* are the only organisms in the database, in which the percentage of homo repeat proteins exceeds that of hetero perfect repeat proteins. In *Guillardia theta* and *Encephalitozoon cuniculi*, the percentage of all types of perfect repeat proteins is almost the same as that of perfect hetero repeat proteins alone, which implies that the number of homo repeat proteins is quite less and in fact lowest amongst the eukaryotes studied here. However, it is warned that several RefSeq genomes are still under review and contain several proteins sequences yet to be verified, and *G. theta* and *E. cuniculi* genomes are examples of such genomes.

Rat and mouse proteins have lower number of perfect repeat proteins as compared to human proteome (Figure 3b). However, percentage of homo repeat proteins in rat and mouse are almost same and only marginally lower than that of humans, implying that the human proteome is richer in proteins with hetero repeats. It might have some implications regarding the observation that the generation of repeats promote protein evolution [44] and formation of novel functional variants [45,46]. We observed differential distribution of repeats in other closely related species as well, for example *Saccharomyces cerevisiae* and *S. pombe* have almost equal fraction of repeat proteins, however, the percentage of homo repeat proteins is considerably higher in *S. cerevisiae* (11%) as compared to *S. pombe* (~7%)(Figure 3b).

The percentage of repeat proteins in prokaryotic genomes is lower as compared to that in eukaryotes. The highest percentage of repeat proteins in prokaryotes is observed in *Thermus thermophilus* (~29%) followed by *Mycobacterium tuberculosis Rv* (~28%) whereas *Salmonella typhi* contains the lowest number of repeat proteins not only amongst prokaryotes, but in all the proteomes studied here (Figure 3c). It is interesting to note that *M. tuberculosis Rv* has higher percentage of repeat proteins as compared to that in other mycobacterium species, namely *M. bovis* and *Mycobacterium avium*. All the mycobacterial species except *Mycobacterium leprae*, predominantly have alanine, glycine and proline homo repeat proteins. Alanine repeats are abundant in PPE (proteins with proline-proline-glutamine motifs) family and a few PE-PGRS proteins (proteins with proline-glutamine motifs and polymorphic GC-rich repetitive sequences) [47] of *M. tuberculosis Rv* and *M. bovis*. However, all *M. avium* alanine homo repeat proteins are hypothetical proteins except two: one PE family protein and a PstA (phosphate-specific transport subunit A) protein. *M. leprae* has the lowest percentage of repeat proteins amongst mycobacterium genomes in the database mainly due to lower numbers of PPE family homologs. The glycine repeats are exclusively present in PE-PGRS proteins of *M. tuberculosis Rv* and *M. bovis* but absent in *M. avium* and *M. leprae*. The PE-PGRS is a newly identified family of fibronectin-binding proteins involved in antigenic variation. The number of glycine repeat proteins in *M. avium* and *M. leprae* is much less as compared to that in the other two species. All the four species of mycobacterium have proline repeats, mainly in the proline-rich antigens. Repetitive regions have been used as genetic markers for the strain differentiation and epidemiology of mycobacterium species [48,49]. It will be interesting to perform systematic comparative investigation of different types of repeats in the species and investigate possible roles of repeats in specific interactions with their hosts.

Amongst the archeal genomes, the percentage of total repeat proteins is highest in *Halobacterium* (~21%) followed by *Aeropyrum pernix* and the lowest in *Archaeoglobus fulgidus* (~7%) (Figure 3d). Notably, we found that all archeal proteomes have lower number of homo repeat proteins. The highest percentage of homo repeat proteins is observed in *Halobacterium* sps. (~4%) and lowest in *A. fulgidus* (~1%). During evolution, archeal genomes mainly accumulated hetero repeat proteins but the resistance towards evolution of homo repeat proteins or mechanism of specific selection of hetero duplication in genomes is not yet known. A detailed and comprehensive analysis of different types of repeats in the three super kingdoms as well as intra-kingdom might provide leads to answers of many open questions, which are, however, beyond the scope of the present study.

The ProtRepeatsDB contains tools for large-scale comparison of protein repeats across genomes to aid studies related to evolution of repeat genesis and functional roles of such repeats in different organisms. The differential propensity of repeats among the super kingdoms/lineages and organisms within the same kingdom/genus emphasizes that different types of repeats have undergone different selection pressure and propagation mechanisms during evolution. However, conclusions derived from analysis of sequences emerging from the sequencing projects with little or no manual curation or additional experimental validation of gene structure have to be dealt with additional care, as some of such sequences may have translations from incorrect gene predictions.

Conclusion

ProtRepeatsDB is a multi-organism database of protein repeats, which is the first database of its kind that incorporates different kinds of repeats viz. perfect repeats-homopeptides and heteropeptides, mismatch repeats and profile patterns representing different families of repeats. The current version (v 1.2) consists of 120686 perfect repeats, 834621 mismatch repeats and 3673 profile repeats from 894890 protein sequences belonging to 141 genomes. The web interface of ProtRepeatsDB consists of unique tools which allow formulation of queries for retrieval and cross species comparison of repeats.

Availability and requirements

ProtRepeatsDB is freely accessible on the Internet at http://bioinfo.icgeb.res.in/repeats/ (Figure 1b). The database is divided into integrated sections of different types of repeats for easy browsing and data retrieval. The web interface of ProtRepeatsDB is supported with PERL and PHP scripts which enable formulation of queries against the database. Results are displayed either in tabulated or graphical formats.

Authors' contributions

Corresponding author- DG, and VSC conceived the project. DG wrote the manuscript, MKK and VSC reviewed it. MKK, SD, GR and DG developed the MySQL database, Web interface and related PHP scripts. GR and DG developed DIREP.

Future directions

ProtRepeatsDB will be regularly updated with protein repeat sequences in emerging annotated sequences from various genome sequencing projects. ProtRepeatsDB will be developed further to include cross links with other databases, repeats detected by other repeat finding algorithms, 3-dimensional structures of repeat proteins, web based repeat finding servers, tools for phylogenetic analysis and ortholog based search for comparative analysis of repeats.

Acknowledgements

We thank WHO (World Health Organization) for the support to the bioinformatics facility at ICGEB. MKK acknowledges CSIR (Council for Scientific and Industrial Research, India) and GR acknowledge UGC (University Grants Commission, India) for financial support. We also thank Arnab Pain (The Wellcome Trust Sanger Institute) for reviewing the manuscript.

References

1. Karlin S, Brocchieri L, Bergman A, Mrazek J, Gentles AJ: **Amino acid runs in eukaryotic proteomes, disease associations.** *Proc Natl Acad Sci USA* 2002, **99:**333-338.
2. Hofferbert S, Schanen NC, Chehab F, Francke U: **Trinucleotide repeats in the human genome:size distributions for all possible triplets, detection of expanded disease alleles in a group of Huntington disease individuals by the repeat expansion detection method.** *Hum Mol Genet* 1997, **6:**77-83.
3. Kazemi-Esfarjani P, Trifiro MA, Pinsky L: **Evidence for a repressive function of the long polyglutamine tract in the human androgen receptor:possible pathogenetic relevance for the CAGn-expanded neuronopathies.** *Hum Mol Genet* 1995, **4:**523-527.
4. Brown LY, Brown SA: **Alanine tracts:the expanding story of human illness, trinucleotide repeats.** *Trends Genet* 2004, **20:**51-58.
5. Perutz MF, Johnson T, Suzuki M, Finch JT: **Glutamine repeats as polar zippers:their possible role in inherited neurodegenerative diseases.** *Proc Natl Acad Sci USA* 1994, **91:**5355-5358.
6. Michelitsch MD, Weissman JS: **A census of glutamine/asparagine-rich regions:implications for their conserved function, the prediction of novel prions.** *Proc Natl Acad Sci USA* 2000, **97:**11910-11915.
7. Stahl N, Prusiner SB: **Prions, prion proteins.** *FASEB J* **5:**2799-2807.
8. Leikauf GD, Borchers MT, Prows DR, Simpson LG: **Mucin apoprotein expression in COPD.** *Chest* 1991, **121:**166S-182S.
9. Clarke JL, Sodeinde O, Mason PJ: **A unique insertion in Plasmodium berghei glucose-6-phosphate dehydrogenase-6-phosphogluconolactonase: evolutionary and functional studies.** *Mol Biochem Parasitol* 2003, **127:**1-8.
10. Reeder JC, Brown GV: **Antigenic variation, immune evasion in Plasmodium falciparum malaria.** *Immunol Cell Biol* 1996, **74:**546-554.
11. Anders RF, Smythe JA: **Polymorphic antigens in Plasmodium falciparum.** *Blood* 1989, **74:**1865-1875.
12. Verra F, Hughes AL: **Biased amino acid composition in repeat regions of Plasmodium antigens.** *Mol Biol Evol* 1999, **16:**627-33.
13. Lee MG, Russell DG, D'Alesandro PA, Van der Ploeg LH: **Identification of membrane-associated proteins in Trypanosoma brucei encoding an internal, EARLRAEE amino acid repeat.** *J Biol Chem* 1994, **269:**8408-8415.
14. Gravekamp C, Rosner B, Madoff LC: **Deletion of repeats in the alpha C protein enhances the pathogenicity of group B streptococci in immune mice.** *Infect Immun* 1998, **66:**4347-4354.
15. Parham SN, Resende CG, Tuite MF: **Oligopeptide repeats in the yeast protein Sup35p stabilize intermolecular prion interactions.** *EMBO J* 2001, **20:**2111-2119.
16. Andrade MA, Perez-Iratxeta C, Ponting CP: **Protein repeats:structures, functions, evolution.** *J Struct Biol* 2001, **134:**117-131.
17. Heger A, Holm L: **Rapid automatic detection, alignment of repeats in protein sequences.** *Proteins* 2000, **41:**224-237.
18. Andrade MA, Ponting CP, Gibson TJ, Bork P: **Homology-based method for identification of protein repeats using statistical significance estimates.** *J Mol Biol* 2000, **298:**521-537.
19. Wan H, Li L, Federhen S, Wootton JC: **Discovering simple regions in biological sequences associated with scoring schemes.** *J Comput Biol* 2003, **10:**171-185.
20. Alba MM, Santibanez-Koref MF, Hancock JM: **The comparative genomics of polyglutamine repeats:extreme differences in the codon organization of repeat-encoding regions between mammals, Drosophila.** *J Mol Evol* 2001, **52:**249-259.
21. Alba MM, Guigo R: **Comparative analysis of amino acid repeats in rodents, humans.** *Genome Res* 2004, **14:**549-554.
22. Depledge DP, Dalby AR: **COPASAAR – A database for proteomic analysis of single amino acid repeats.** *BMC Bioinformatics* 2005, **6:**196.

23. **NCBI Reference sequences (RefSeq)** [http://www.ncbi.nlm.nih.gov/RefSeq/]
24. Karlin S: **Statistical significance of sequence patterns in proteins.** *Curr Opin Struct Biol* 1995, **5**:360-371.
25. Mott R: **Accurate formula for P-values of gapped local sequence, profile alignments.** *J Mol Biol* 2000, **300**:649-659.
26. Hulo N, Sigrist CJ, Le Saux V, Langendijk-Genevaux PS, Bordoli L, Gattiker A, De Castro E, Bucher P, Bairoch A: **Recent improvements to the PROSITE database.** *Nucleic Acids Res* 2004:D134-D137.
27. Rice P, Longden I, Bleasby A: **EMBOSS:the European Molecular Biology Open Software Suite.** *Trends Genet* 2000, **16**:276-277.
28. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Res* 1997, **25**:3389-3402.
29. Brunori M, Silvestrini MC, Pocchiari M: **The scrapie agent, the prion hypothesis.** *Trends Biochem Sci* 1988, **13**:309-313.
30. Kuznetsov IB, Morozov PS, Matushkin YG: **Prion proteins:evolution, preservation of secondary structure.** *FEBS Lett* 1997, **412**:429-432.
31. Croes EA, Theuns J, Houwing-Duistermaat JJ, Dermaut B, Sleegers K, Roks G, Van den Broeck M, van Harten B, van Swieten JC, Cruts M, Van Broeckhoven C, van Duijn CM: **Octapeptide repeat insertions in the prion protein gene, early onset dementia.** *J Neurol Neurosurg Psychiatry* 2004, **75**:1166-1170.
32. Ferlin A, Garolla A, Bettella A, Bartoloni L, Vinanzi C, Roverato A, Foresta C: **Androgen receptor gene CAG, GGC repeat lengths in cryptorchidism.** *Eur J Endocrinol* 2005, **152**:419-425.
33. Ding D, Xu L, Menon M, Reddy GP, Barrack ER: **Effect of GGC glycine repeat length polymorphism in the human androgen receptor on androgen action.** *Prostate* 2005, **62**:133-139.
34. Das AK, Cohen PW, Barford D: **The structure of the tetratricopeptide repeats of protein phosphatase 5:implications for TPR-mediated protein-protein interactions.** *EMBO J* 1998, **17**:1192-1199.
35. Goebl M, Yanagida M: **The TPR snap helix:a novel protein repeat motif from mitosis to transcription.** *Trends Biochem Sci* 1991, **16**:173-177.
36. D'Andrea LD, Regan L: **TPR proteins:the versatile helix.** *Trends Biochem Sci* 2003, **28**:655-662.
37. Main ER, Stott K, Jackson SE, Regan L: **Local and long-range stability in tandemly arrayed tetratricopeptide repeats.** *Proc Natl Acad Sci USA* 2005, **102**:5721-5726.
38. Tanaka M, Machida Y, Nishikawa Y, Akagi T, Hashikawa T, Fujisawa T, Nukina N: **Expansion of polyglutamine induces the formation of quasi-aggregate in the early stage of protein fibrillization.** *J Biol Chem* 2003, **278**:34717-34724.
39. de Pril R, Fischer DF, Maat-Schieman ML, Hobo B, de Vos RA, Brunt ER, Hol EM, Roos RA, van Leeuwen FW: **Accumulation of aberrant ubiquitin induces aggregate formation and cell death in polyglutamine diseases.** *Hum Mol Genet* 2004, **13**:1803-1813.
40. Singh GP, Chandra BR, Bhattacharya A, Akhouri RR, Singh SK, Sharma A: **Hyper-expansion of asparagines correlates with an abundance of proteins with prion-like domains in Plasmodium falciparum.** *Mol Biochem Parasitol* 2004, **137**:307-19.
41. Zeller R, Haramis AG, Zuniga A, McGuigan C, Dono R, Davidson G, Chabanis S, Gibson T: **Formin defines a large family of morphoregulatory genes and functions in establishment of the polarising region.** *Cell Tissue Res* 1999, **296**:85-93.
42. Pires JR, Parthier C, Aido-Machado R, Wiedemann U, Otte L, Bohm G, Rudolph R, Oschkinat H: **Structural basis for APPTPPPLPP peptide recognition by the FBP11WW1 domain.** *J Mol Biol* 2005, **348**:399-408.
43. Chiang PK, Bujnicki JM, Su X, Lanar DE: **Malaria: therapy, genes and vaccines.** *Curr Mol Med* 2006, **6**:309-26.
44. Tompa P: **Intrinsically unstructured proteins evolve by repeat expansion.** *Bioessays* 2003, **25**:847-55.
45. Marcotte EM, Pellegrini M, Yeates TO, Eisenberg D: **A census of protein repeats.** *J Mol Biol* 1999, **293**:151-60.
46. Williamson MP: **The structure and function of proline-rich regions in proteins.** *Biochem J* 1994, **297**:249-60.
47. Espitia C, Laclette JP, Mondragon-Palomino M, Amador A, Campuzano J, Martens A, Singh M, Cicero R, Zhang Y, Moreno C: **The PE-PGRS glycine-rich proteins of Mycobacterium tuberculosis: a new family of fibronectin-binding proteins?** *Microbiology* 1999, **145**:3487-3495.
48. Sun YJ, Lee AS, Ng ST, Ravindran S, Kremer K, Bellamy R, Wong SY, van Soolingen D, Supply P, Paton NI: **Characterization of ancestral Mycobacterium tuberculosis by multiple genetic markers and proposal of genotyping strategy.** *J Clin Microbiol* 2004, **42**:5058-5064.
49. van Soolingen D, de Haas PE, Hermans PW, Groenen PM, van Embden JD: **Comparison of various repetitive DNA elements as genetic markers for strain differentiation and epidemiology of Mycobacterium tuberculosis.** *J Clin Microbiol* 1993, **31**:1987-1995.