International Journal for Computational Biology (IJCB)
Vol.4, No.1, April 2015, pp. 62~66
ISSN: 2278-8115

DPAAR: a Database of Perfect Amino Acid Repeat

Himansu Kumar [1], Swati Srivastava [2], Pritish Kumar Varadwaj [1*]
[1] Indian Institute of Information Technology Allahabad, India
[2] Lovely Professional Universitiy, Jalandhar, Punjab, India

Article Info

Article history:

Received Jul 22nd, 2014
Revised Nov 20th, 2014
Accepted Dec 24th, 2014

Keyword:

Swiss-Prot database,
DPAAR, repeate sequence.

ABSTRACT

Repeat of amino acids in a protein sequence has clinical and functional importance. Database of Perfect Amino Acid Repeat (DPAAR) is a kind of relational as well as flat file database which is created by the comprehensive analysis of 5,42,782 protein sequences of Swiss-Prot database (released on 19th March,2014) to know the association between repeated sequence and disease. It provides the search engine for rapid access of a particular repeated amino acid, or particular swissprot ID, or particular length of repeated amino acids in a protein sequence. It also provides the flat files for single, oligo, and tandem repeated sequence information to get the complete informaton about concerned amino acids repeat. It consists of the tables of repeated sequence and its associated disease in human being.

Corresponding Author:

Pritish Kumar Varadawj
Indian Institute of Information
Technology Allahabad.
Email: pritish@iiita.ac.in

How to Cite:

Himansu Kumar *et. al.* DPAAR: a database of Perfect Amino Acid Repeat. IJCB. 2015; Volume 4 (Issue 1): Page 62-66.

1. INTRODUCTION

Recent findings approves that the repeats of single, oligo, and tandem amino acid in a protein sequence is playing crucial role in various functional and evolutionary aspects ,specially its close proximity with various disease like neurodegenerative disorder, cancer, muscular dystrophy etc. Amino acid repeat can be perfect repeat or a mismatch repeat; repeats can be further of few amino acids to long span of repeat [1].

Repeats can be classified as Homopeptide repeat or Monopeptide repeat , containing same amino acid repeat to a stretch and Heteropeptide repeat (including oligopeptide and periodic repeats) containing amino acid repeat with some combination of other amino acid repeat [2].

The occurrence of repeated amino acid in a protein sequence is surprisingly distributed in a heterogeneous manner like presence of glycine and glutamine repeat is very high whereas presence of tryptophan repeat is negligible in whole swiss-prot database. Recent findings approves that approx 14% of all proteins containing the internal repeats and occurrence of repeats in eukaryotic protein is higher than prokaryotic protein and repetition of glutamine (30-40) has been reported in various neural diseases [3]. Many online databases efficiently describing the amino acid repeats, such as Tandem Repeat in a Protein Sequence (TRIPS) is an exclusively flat file data base [4]. ProtRepeat is a relational database of 141 organisms [5] whereas COPASAAR **[6]** is for 244 organisms and exclusively for single amino acid repeats.A database called RepSeq is exclusively for lower eukaryotic pathogens **[7]**.A common platform for search engine, flat files contents as well as repeated sequence and associated disease tables are required. In this work Search engine is designed for rapid access of the database and flat file is for to reduce the run time complexity.

2. RESEARCH METHODOLOGY

2.1 Design and Implementation:

DPAAR database has no sequence length limitations and can find repeat in sequence ranging from few amino acids to thousands of amino acid length sequence. The database is constructed by the use of MySQL, PHP and PERL scripting languages. It has categorized the repeat under four categories: Single Amino Acid Repeat (SAAR), Oligo Amino Acid Repeat (OAAR), and Periodically Conserved Amino Acid Repeats (PCAA).

For the detection of repeat in all the SwissProt databse, sequences were downloaded from http://www.uniprot.org/downloads. A PERL program was written with the use of regular expression and sliding window method for finding the particular repeat with defined range of there occurrence (e.g.: “A” between 5-10 amino acid repeats) . Sequences were fetched from SwissProt as an array for all protein and input them as string of length “l”. A minimum required repeat range (min) and upper limit (max) was selected, and with the help of regular expression (REGEX) the repeats were obtained. In case of single amino acid and oligo amino acid repeat same method as above was applied. For oligo amino acid detection, amino acid with different permutation and combination were selected and searched.

In case of periodically conserved amino acid, same algorithm as SAAR and OAAR was used except in this a(i) is checked with a(i+2) till ‘n’, where ‘n’ is the minimum value of repeat which is set of 5 for this database. The output obtained was stored in flat file and displayed in HTML pages .The data was also inserted in MySQL for user query.

2.2 Search Engine Design:

Database provides user interface, where the user can input either SwissProt ID of desired protein or monopeptide and oligopeptide amino acid or can fetch protein information based on the required amino acid length, depicted in Figure1. The database search engine displays all protein containing the required search which consists SwissProt Id, Accession No, name of the Amino Acid, Protein name, times or number of repeats, Amino acid length, and repeated sequence as depicted in Figure 2. It is also helpful for comparative study of their structure and sequence similarity between various organisms which are having same amino acid repeat with same amino acid length in other words conserved sequence of repeat can be analyzed in different organism.

2.3 Flat File Design:

The database as mentioned above was categorized like: (a) Single Amino Acid Repeat which contains information for 20 amino acids which is further divided into three – sequence containing 5-10 repeats,10-20 amino acid repeat and more than 20 repeats. (b) Oligo amino acid classified under di, tri, tetra, penta and hexapeptide repeat, depicted in Figure3. (c) Periodically conserved repeat contain information for all 20 amino acid periodically repeated at every consecutive position. (d) Lastly repeat and its associated repeat contains information of disease [8] caused by mono, di and tripeptiderepeats, depicted in Figure4.

2.4 Current database summary:

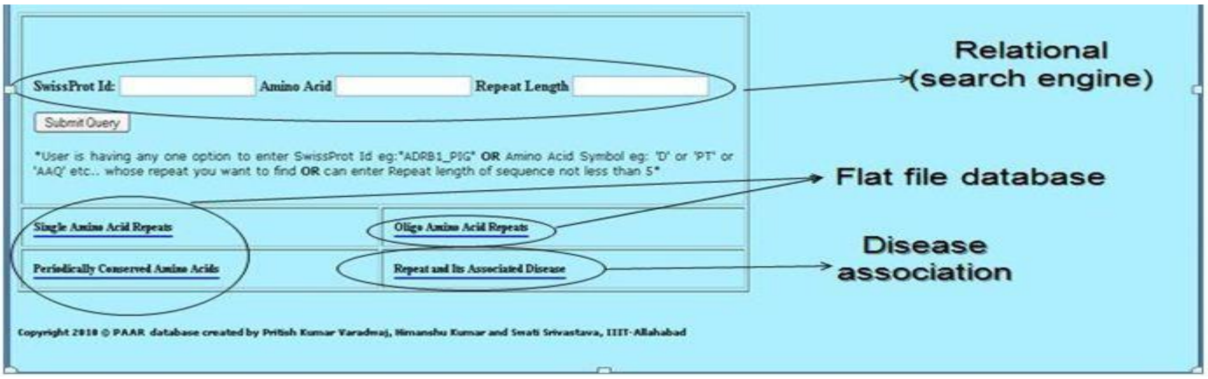
Out of complete protein sequences we have detected amino acid repeats in 49,400 protein sequences. Availability of the glutamine, proline and aspartic acid is very high where as presence of tryptophan were very low depicted as in Table1 and Figure 4. It has been observed that the presence of glutamine more than 35 is alarming, causing lots of neural diseases. As per present survey mainly alanine, glutamine and glycine are involved in disease causing situation**.**

2.5 Dataset Generation:

Datasets for the DPAAR database were extracted from SWISSPROT database.
Source: ftp://ftp.uniprot.org/pub/databases/uniprot/

2.6 Current Location

Web Link: http://maahanswahini.com/DPAAR/



SwissProt Id: Amino Acid Repeat Length

Submit Query

*User is having any one option to enter SwissProt Id eg:"ADRB1_PIG" **OR** Amino Acid Symbol eg: 'D' or 'PT' or 'AAQ' etc.. whose repeat you want to find **OR** can enter Repeat length of sequence not less than 5*

Single Amino Acid Repeats

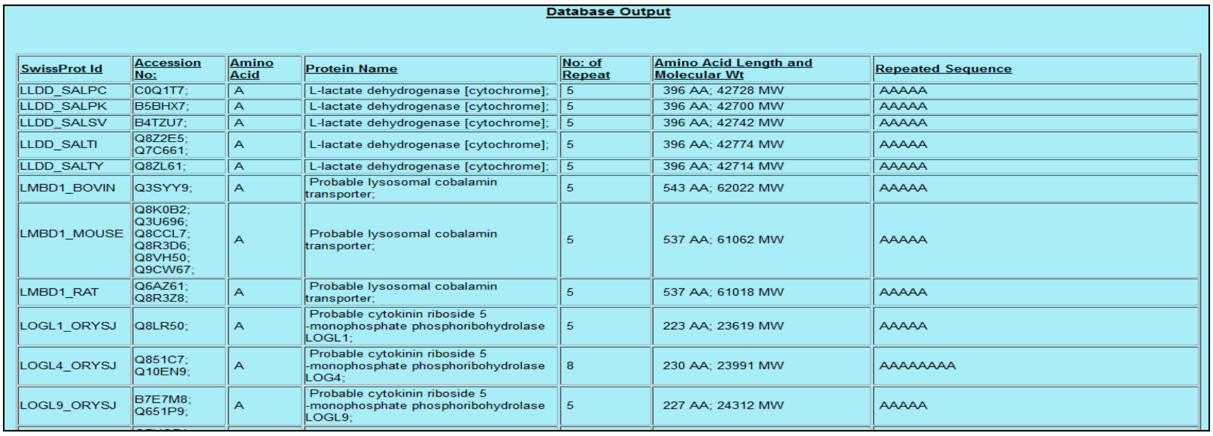
Oligo Amino Acid Repeats

Periodically Conserved Amino Acids

Repeat and Its Associated Disease

Copyright 2010 © PAAR database created by Pritish Kumar Varadwaj, Himanshu Kumar and Swati Srivastava, IIIT-Allahabad

Figure 1: Searching options of the database.



Database Output

SwissProt Id	Accession No:	Amino Acid	Protein Name	No: of Repeat	Amino Acid Length and Molecular Wt	Repeated Sequence
LLDD_SALPC	C0Q1T7;	A	L-lactate dehydrogenase [cytochrome];	5	396 AA; 42728 MW	AAAAA
LLDD_SALPK	B5BHX7;	A	L-lactate dehydrogenase [cytochrome];	5	396 AA; 42700 MW	AAAAA
LLDD_SALSV	B4TZU7;	A	L-lactate dehydrogenase [cytochrome];	5	396 AA; 42742 MW	AAAAA

LLDD_SALTI	Q8Z2E5;
	Q7C661;
		A	L-lactate dehydrogenase [cytochrome];	5	396 AA; 42774 MW	AAAAA

LLDD_SALTY	Q8ZL61;	A	L-lactate dehydrogenase [cytochrome];	5	396 AA; 42714 MW	AAAAA
LMBD1_BOVIN	Q3SYY9;	A	Probable lysosomal cobalamin transporter;	5	543 AA; 62022 MW	AAAAA

LMBD1_MOUSE	Q8K0B2;
	Q3U696;
	Q8CCL7;
	Q8R3D6;
	Q8VH50;
	Q9CW67;
		A	Probable lysosomal cobalamin transporter;	5	537 AA; 61062 MW	AAAAA

LMBD1_RAT	Q6AZ61;
	Q8R3Z8;
		A	Probable lysosomal cobalamin transporter;	5	537 AA; 61018 MW	AAAAA

LOGL1_ORYSJ	Q8LR50;	A	Probable cytokinin riboside 5 -monophosphate phosphoribohydrolase LOGL1;	5	223 AA; 23619 MW	AAAAA

LOGL4_ORYSJ	Q851C7;
	Q10EN9;
		A	Probable cytokinin riboside 5 -monophosphate phosphoribohydrolase LOG4;	8	230 AA; 23991 MW	AAAAAAAA

LOGL9_ORYSJ	B7E7M8;
	Q651P9;
		A	Probable cytokinin riboside 5 -monophosphate phosphoribohydrolase LOGL9;	5	227 AA; 24312 MW	AAAAA

Figure 2: Database output display.

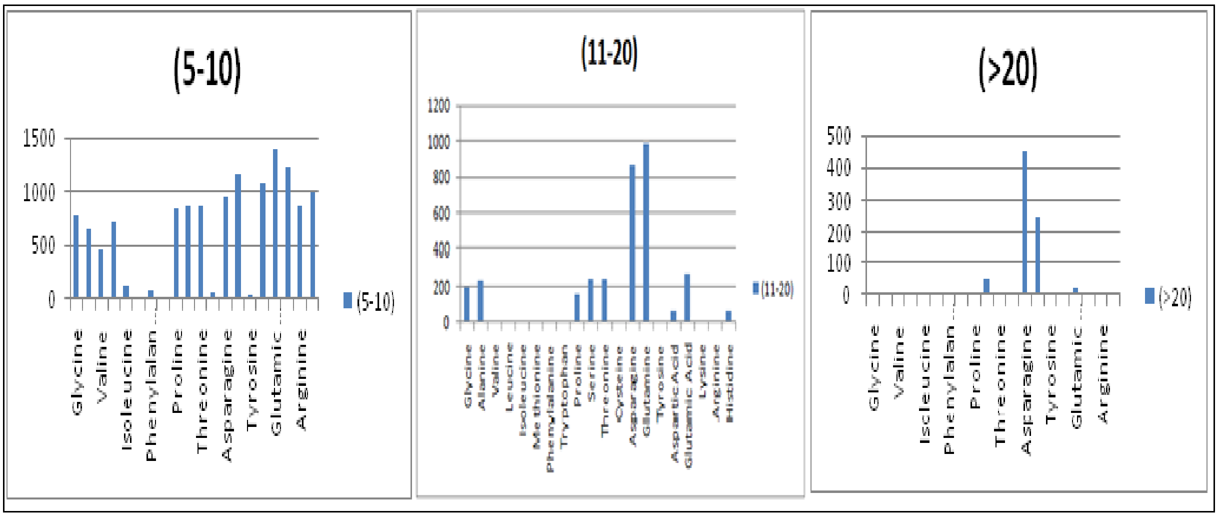
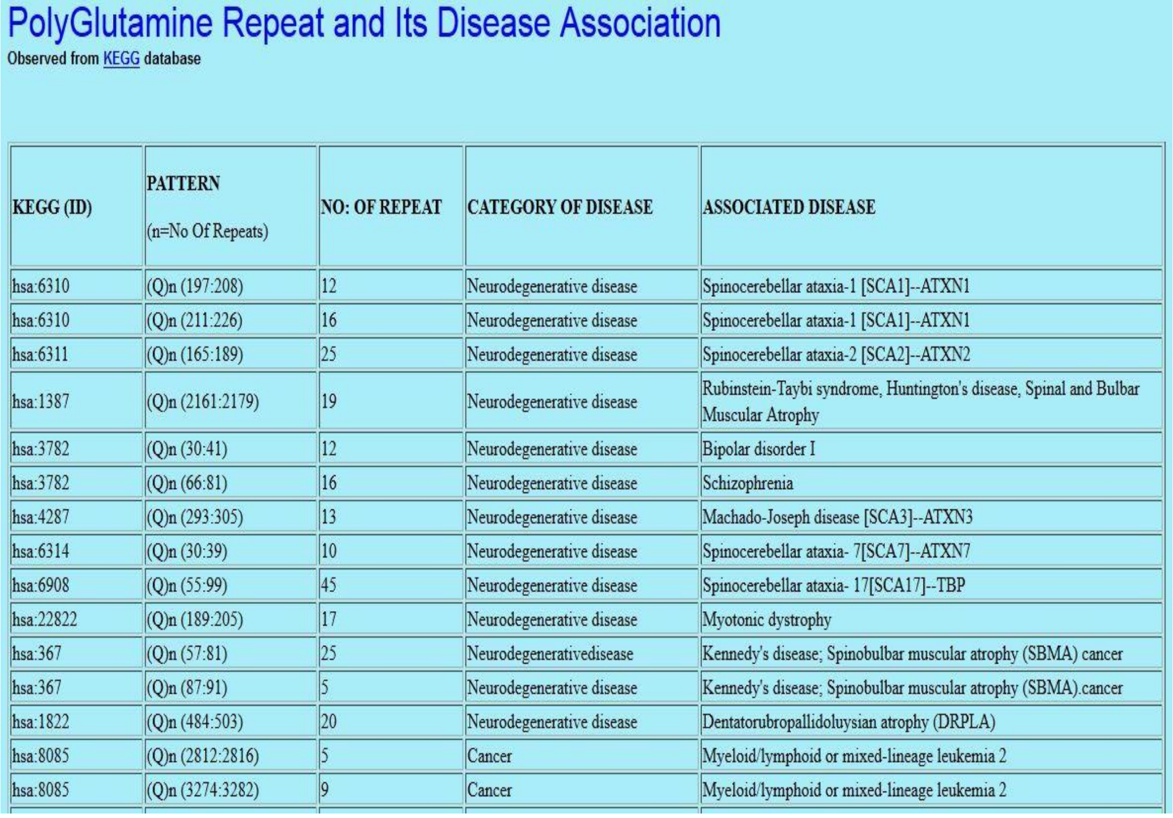


Figure 3(a): Repeatations of the aminoacids between 5-10, (b) Repeatations of the aminoacids between 11-20, (c) Repeatations of the aminoacids >20 in the whole swissprot database.

Amino acids	***5-10***	***11-20***	***>20***
Glycine	790	200	7
Alanine	650	234	1
Valine	468	1	No
Leucine	722	14	No
Isoleucine	118	1	1
Methionine	27	no	No
Phenylalanine	92	1	No
Tryptophan	1	No	No
Proline	850	160	9
Serine	870	240	53
Threonine	877	238	3
Cysteine	69	1	No
Asparagine	943	870	450
Glutamine	1165	987	250
Tyrosine	33	1	No
Aspartic Acid	1070	68	9
Glutamic Acid	1400	270	23
Lysine	1233	4	No
Arginine	876	5	No
Histidine	986	62	2

Table 1: Overall analysis of the database.



PolyGlutamine Repeat and Its Disease Association

Observed from KEGG database

KEGG (ID)	PATTERN (n=No Of Repeats)	NO: OF REPEAT	CATEGORY OF DISEASE	ASSOCIATED DISEASE
hsa:6310	(Q)n (197:208)	12	Neurodegenerative disease	Spinocerebellar ataxia-1 [SCA1]--ATXN1
hsa:6310	(Q)n (211:226)	16	Neurodegenerative disease	Spinocerebellar ataxia-1 [SCA1]--ATXN1
hsa:6311	(Q)n (165:189)	25	Neurodegenerative disease	Spinocerebellar ataxia-2 [SCA2]--ATXN2
hsa:1387	(Q)n (2161:2179)	19	Neurodegenerative disease	Rubinstein-Taybi syndrome, Huntington's disease, Spinal and Bulbar Muscular Atrophy
hsa:3782	(Q)n (30:41)	12	Neurodegenerative disease	Bipolar disorder I
hsa:3782	(Q)n (66:81)	16	Neurodegenerative disease	Schizophrenia
hsa:4287	(Q)n (293:305)	13	Neurodegenerative disease	Machado-Joseph disease [SCA3]--ATXN3
hsa:6314	(Q)n (30:39)	10	Neurodegenerative disease	Spinocerebellar ataxia- 7[SCA7]--ATXN7
hsa:6908	(Q)n (55:99)	45	Neurodegenerative disease	Spinocerebellar ataxia- 17[SCA17]--TBP
hsa:22822	(Q)n (189:205)	17	Neurodegenerative disease	Myotonic dystrophy
hsa:367	(Q)n (57:81)	25	Neurodegenerativedisease	Kennedy's disease; Spinobulbar muscular atrophy (SBMA) cancer
hsa:367	(Q)n (87:91)	5	Neurodegenerative disease	Kennedy's disease; Spinobulbar muscular atrophy (SBMA).cancer
hsa:1822	(Q)n (484:503)	20	Neurodegenerative disease	Dentatorubropallidoluysian atrophy (DRPLA)
hsa:8085	(Q)n (2812:2816)	5	Cancer	Myeloid/lymphoid or mixed-lineage leukemia 2
hsa:8085	(Q)n (3274:3282)	9	Cancer	Myeloid/lymphoid or mixed-lineage leukemia 2

Figure 4: Association of the disease with repeated amino acid sequences.

4. ACKNOWLEDGEMENTS

The Author would like to thank Director of Indian Institute of information Technology, Allahabad, India for providing valuable infrastructure and research environment to complete this work.

5. REFERENCES

[1] Miguel A. Andrade,Carolina Perez-Iratxeta,Chris P. Ponting. Protein Repeats: Structures, Functions, and Evolution. J Struct Biol. 2001 May-Jun; 134(2-3):117-31.

[2] Szklarczyk R, Heringa J, Tracking repeats using significance and transitivity, Bioinformatics, 2004 Aug 4; 20 Suppl :i311-7.

[3] Marcotte EM, Pellegrini M, Yeates TO, Eisenberg D. A census of protein repeats. J MolBiol, 1999, 293:151-160.

[4] Katti MV, Sami-Subbu R, Ranjekar PK, and Gupta VS: Amino acid repeat patterns in protein sequences: Their diversity and structural-functional implications.Protein Sci 2000, 9:1203-1209.

[5] Kalita MK, Ramasamy G, Duraisamy S, and Chauhan VS, Gupta D: ProtRepeatsDB: A database of amino acid repeats in genomes. BMC Bioinformatics, 2006, 7(1):336.

[6] Depledge DP, Dalby AR, COPASAAR-A database for proteomic analysis of single amino acid repeats, BMC Bioinformatics 2005, 6:196.

[7] Daniel P Depledge, Ryan PJ Lower and Deborah F Smith, RepSeq – A database of amino acid repeats present in lower eukaryotic pathogens. BMC Bioinformatics 2007, 8:122.

[8] Karlin S, Bracchieri L, Bergman A, and Mrazek J, Gentles AJ: Amino acid runs in eukaryotic proteomes, disease associations.ProcNatlAcadSciUSA2002, 99: 333-338.