BIOINFORMATION
Discovery at the interface of physical and biological sciences

open access www.bioinformation.net Software

Volume 10(1)

PPS: A computing engine to find Palindromes in all Protein sequences

Zameer Ahmed, Manickam Gurusaran, Prasanth Narayana, Kala Sekar Dinesh Kumar, Jayapal Mohanapriya, Marthandan Kirti Vaishnavi & Kanagaraj Sekar*

Supercomputer Education and Research Centre, Indian Institute of Science, Bangalore 560 012, India; Kanagaraj Sekar - Email: sekar@physics.iisc.ernet.in; Phone: +91-080-22933059/23600551; Fax: +91-080-23600683; *Corresponding author

Received January 05, 2014; Revised January 23, 2014; Accepted January 24, 2014; Published January 29, 2014

Abstract:
The primary structure of a protein molecule comprises a linear chain of amino acid residues. Certain parts of this linear chain are unique in nature and function. They can be classified under different categories and their roles studied in detail. Two such unique categories are the palindromic sequences and the Single Amino Acid Repeats (SAARs), which plays a major role in the structure, function and evolution of the protein molecule. In spite of their presence in various protein sequences, palindromes have not yet been investigated in detail. Thus, to enable a comprehensive understanding of these sequences, a computing engine, PPS, has been developed. The users can search the occurrences of palindromes and SAARs in all the protein sequences available in various databases and can view the three-dimensional structures (in case it is available in the known three-dimensional protein structures deposited to the Protein Data Bank) using the graphics plug-in Jmol. The proposed server is the first of its kind and can be freely accessed through the World Wide Web.

Availability: URL http://pranag.physics.iisc.ernet.in/pps/.

Keywords: Single Amino Acid Repeats, computing engine, three-dimensional crystal structures, Palindromes.

Background:
The three-dimensional structure of a protein molecule is determined by its amino acid sequence, which is known to display many distinctive patterns often repeating itself. In protein sequences, these repeats can be catalogued based on their orientation, such as – Direct and inverted repeats **[1];** that are known to play important roles in protein structure and function **[2, 3, 4].** Palindromic sequence (a special type of inverted repeats with no spacer sequence) is a symmetric set of characters that looks the same when read from two directions (forward and backward). Palindromes in DNA and RNA are mostly linked with restriction enzyme recognition sites **[5],** termination of RNA polymerase III **[6]** and in the formation of hairpin loops. While palindromes in proteins are observed in different classes of proteins like histones, prion proteins **[7, 8, 9]**, DNA binding proteins **[2, 4, 10],** Rhodopsin family **[11],** metal binding proteins **[12],** sugar metabolizing proteins **[13]** and receptors **[14]** carrying out complex functions. However, the precise function of these palindromic sequences has not yet been fully understood. Palindromic sequences in proteins were identified two decades ago and since then, it has been shown that they are mainly seen in low complexity regions and display a high tendency to form helices **[15].** Apart from this, not much information is available on these palindromes. Secondly, SAARs (Single Amino Acid Repeats) also known as "Homopeptides" (a special type of direct repeats with no spacer sequence) are repeats that are formed by a single amino acid being repeated in tandem and are believed to arise due to

the sequential expansion of short codons in the genome. These repeats have also often associated with regulatory network of epigenesis, replication, transcription and evolution **[3, 16].** In addition, they are linked to the development and activity of a peripheral nervous system **[17]** and with diseases like Huntington disease and other nervous system related diseases. In view of the above, a computing engine, PPS (Palindromes in Protein Sequences) has been developed to identify palindrome and SAARs in the protein sequences available in various databases. To the best knowledge of the authors, there exists no efficient and robust web engine. Thus, the development of a computing engine for the identification of these sequences will enable biologists and bioinformaticians better to understand palindromes and SAARs present in various protein sequences.

Figure 1: (A) The spatial conformation of "LTIITL" in the crystal structure of a p53 tumor suppressor-DNA complex; **(B)** The inset depicts the three-dimensional structure of the one of the SAARs sequences "GGGGGG" (PDB-id: 2X4M); **(C)** The inset shows the three-dimensional structure of the longest palindromic sequence "TGAKALAKAGT" (PDB-id: 2GOK).

Methodology:

Our algorithm **[18]** employs a strategic two step procedure to extract the palindromes. Firstly, it finds all possible palindrome (including odd and even length palindromes) and next efficiently discards all unnecessary sub-palindromes, to avoid redundancy. PPS is an efficient computing engine that makes use of the above cited algorithm to identify palindromic peptide sequences and SAARs. The user can use the computing engine to identify palindromes in any of the three following ways: **(a)** The amino acid sequence of a known protein sequence can be entered in the text box provided (or); **(b)** The FASTA sequence file can be directly uploaded (or); **(c)** To analyze all the palindrome and SAAR sequences, an option has been provided for the users to choose a particular database from the available sequence databases such as SWISSPROT **[19],** GDB (In-house Genome database), PIR **[20]** and structure databases like PDB **[21],** 25% and 90% non-redundant databases **[22].**

It is to be noted that the user can simultaneously perform all the above three calculations and view the results dynamically. In addition, the users need to enter a value corresponding to the "number of residues in the sequences" option and choose either palindromic sequence or SAARs in the "Show only" option. Based on the above, the resulting window displays either the palindromic sequence or the SAARs details and provides additional options for the users to explore the occurrences of a palindrome or SAAR in other sequence (PIR, SWISS-PROT and GDB) and structural databases (Protein Data Bank [PDB]). As an enhanced feature, the users can also view the three-dimensional structure (in case it is available in the three-dimensional crystal structures deposited in the Protein Data Bank) of the identified palindrome or SAAR using an interactive graphics JAVA plug-in Jmol, interfaced to the proposed computing engine.

Implementation:

The proposed computing engine, PPS, has been developed and optimized using Solaris and is driven by a 2.66 GHz Xeon (R) processor equipped with 4 GB of Random Access Memory (RAM). This operating system was chosen for its reliability and security. The computing engine has been tested vigorously using all platforms (Windows, Linux, Mac and SGI). In general, the computations are very fast and the results are displayed in a rapid time. However, it may vary depending upon the speed and traffic of the network. The computing engine was coded using PERL, JavaScript and HTML in order to develop, validate and design the web pages, respectively. The computing code used for the identification of the palindromes and SAARs have been written using C++. The proposed engine is freely available for all academic users and non-commercial organizations over the World Wide Web at: http: //pranag.physics.iisc.ernet.in /pps/.

Case study:

Palindromes in Tumor Suppressor Proteins

Palindromes are mostly abundant in DNA binding proteins **[4].** p53 is a DNA-binding, tumor suppressor protein that adopts an

intricate mode of action in the reparation through various regulatory mechanisms. Mutations in p53 are associated with human cancers. It also prompts apoptosis when the DNA is irreparable. PPS is used to investigate the p53 protein (GenBank accession No. NP_000537; length: 393 amino acids; Organism: Homo sapiens) for palindromes of length greater than 5. PPS identifies six palindromes **Table 1 (see supplementary material)** in 0.13 seconds, one of which (LTIITL) is known to characterize the DNA binding domain of the protein **[23].** The spatial conformation of "LTIITL" in the crystal structure of a p53 tumor suppressor-DNA complex (PDB-id: 1TSR) is displayed in **(Figure 1A)**. Upon further analysis (results not shown), it is observed that the palindrome "LTIITL" is a key part of the p53 DNA-binding domain. PPS can process multiple sequences, which is a noteworthy and a useful feature for researchers to study large databases of proteins.

Further Case Studies:

Four SAARs (of lengths 6, 12, 6 and 7) are identified (results not shown), when the amino acid sequence of formin-like protein (Uniprot accession No. Q9XIE0; length: 929 amino acids; Organism: *Arabidopsis thaliana*) is uploaded as the input sequence. The number of residues is set to be greater than five residues. **Figure 1B** shows the three-dimensional structure (highlighted) of one of the SAAR's sequences "GGGGGG" observed in the crystal structure (PDB-id: 2X4M) of the plasminogen activator PLA from *Yersinia* pestis.

The amino acid sequence of imidazole nepropionase (Uniprot accession No. Q8U8Z6; length: 419 amino acids; Organism: *Agrobacterium tumefaciens* (strain C58 / ATCC 33970)) is given as the input sequence and here again the "Number of amino acid residues in a palindromic sequence" is set to be greater than five residues. The computing engine identified three palindromic sequences GTATG, MEFEM and TGAKALAKAGT of lengths 5, 5 and 11, respectively (results not shown). The three-dimensional structure (highlighted) of the longest observed palindrome "TGAKALAKAGT" in the crystal structure of the imidazolonepropionase from *Agrobacterium tumefaciens* (PDB-id: 2GOK) is shown in **(Figure 1C).** Interestingly, upon further analysis, it is seen that this fragment is a part of the amidohydro_3 domain, which is responsible for the hydrolysis of amide or amine bonds.

Conclusion:

Palindromes and SAARs are of momentous importance to better understand the structure, function and evolution of proteins. Thus, identifying them will aid biologists and bioinformaticians to understand their significance. Thus, an avant-garde internet computing engine, PPS, has been made available over the World Wide Web to identify palindromes and SAARs. General comments and suggestions for improvements are welcome and should be addressed to Professor K. Sekar at sekar@physics.iisc.ernet.in or sekar@serc.iisc.ernet.in.

Acknowledgement:

The authors gratefully acknowledge the facilities offered by the Supercomputer Education and Research Centre (SERC) and the Interactive graphics facility. One of the authors (KS) thanks the Department of Information Technology (DIT) for financial support in the form of a research grant.

References:

[1] Gurusaran M *et al. Genomics* 2013 **102**: 403 [PMID: 23880222]
[2] Giel-Pietraszuk M *et al. J Protein Chem.* 2003 **22**: 109 [PMID: 12760415]
[3] Uthayakumar M *et al. Genomics Proteomics Bioinformatics* 2012 **10**: 217 [PMID: 23084777]
[4] Ohno S, *Leukemia* 1993 **7**: S157 [PMID: 8361224].
[5] Roulland-Dussoix D & Boyer HW, *Biochim Biophys Acta.* 1969 **195**: 219 [PMID: 4901831].
[6] Chu WM *et al. Nucleic Acids Res.* 1997 **25**: 2077 [PMID: 9153305]
[7] Cheng GH *et al. Proc Natl Acad Sci USA* 1989 **86:** 7002 [PMID: 2780558]
[8] Kazim AL, *FEBS Lett***.** 1993 **331**: 1 [PMID: 8405384]
[9] Sulkowski E, *FASEB J.* 1992 **6:** 2363 [PMID: 1544547]
[10] Suzuki M, *Proc Natl Acad Sci USA.* 1992 **89**: 8726 [PMID: 1528886]
[11] Ohno S, *Riv Biol.* 1990 **83**: 287 [PMID: 2128128]
[12] Pan PK *et al. Eur J Biochem.* 1999 **266**: 33 [PMID: 10542048]
[13] Ohno S, *Hum Genet.* 1992 **90:** 342 [PMID: 1483687]
[14] Jaseja M *et al. J Pept Res.* 2005 **66**: 9 [PMID: 15946191]
[15] Sheari A *et al. BMC Bioinformatics* 2008 **9:** 274 [PMID: 18547401]
[16] Grechko W, *Mol Biol (Mosk)* 2011 **45:** 765 [PMID: 22393774].
[17] Karlin S *et al. Proc Natl Acad Sci USA* 2012 **99**: 333 [PMID: 11782551]
[18] Prasanth N *et al. J Bio Sci.* 2013 **38**: 173 [PMID: 23385825]
[19] Bairoch A & Apweiler R, *Nucleic Acids Res.* 1998 **26**: 38 [PMID: 9399796]
[20] Barker WC *et al. Nucleic Acids Res.* 1998 **26**: 27 [PMID: 9399794].
[21] Berman HM *et al. Nucleic Acids Res.* 2000 **28**: 235 [PMID: 10592235]
[22] Wang G & Dunbrack RL Jr, *Bioinformatics* 2003 **19**: 1589 [PMID: 12912846]
[23] Cho Y *et al. Science* 1994 **265**: 346 [PMID: 8023157]

Edited by P Kangueane

Citation: **Ahmed *et al.*** Bioinformation 10(1): 048-051 (2014)

BIOINFORMATION

open access

Supplementary material:

Table 1: Table displaying palindromes in human cellular tumor antigen p53 isoform A.

Palindrome	Position		Length
	From	To	
PAPAP	85	89	5
PKKKP	318	322	5
SPLPS	33	37	5
LTIITL	252	257	6
EDPGPDE	56	62	7
APAPAAPTPAAPAPA	74	88	15