Bioinformation
www.bioinformation.net

open access
Web Server

FAIR: A server for internal sequence repeats

Ramaswamy Senthilkumar, Radhakrishnan Sabarinathan, Bazil Shaahul Hameed, Nirjhar Banerjee, Narayanan Chidambarathanu, Rajadurai Karthik, Kanagaraj Sekar*

Bioinformatics Centre, Centre of Excellence in Structural Biology and Bio-computing, Indian Institute of Science, Bangalore 560 012, India; K. Sekar – Phone E-mail: sekar@physics.iisc.ernet.in; +91-080-22933059/23602469/23601409; Fax: +91-080-23600683/23600551;
*Corresponding author

Received October 14, 2009; Revised November 16, 2009; Accepted November 16, 2009; Published January 07, 2010

Abstract:
An Internet computing server has been developed to identify all the occurrences of the internal sequence repeats in a protein and DNA sequences. Further, an option is provided for the users to check the occurrence(s) of the resultant sequence repeats in the other sequence and structure (Protein Data Bank) databases. The databases deployed in the proposed computing engine are up-to-date and thus the users will get the latest information available in the respective databases. The server is freely accessible over the World Wide Web (WWW).

Availability: http://bioserver1.physics.iisc.ernet.in/fair/

Keywords: internal repeats, sequence, server, DNA, protein

Background:
With the advent of complete genome sequencing of many organisms, the volume of biological data has been increasing rapidly. Bioinformatics has hence emerged as an indispensable field to organize and assemble the plethora of biological data into a more comprehensive state. For instance, internal sequence repeats (without any gaps) are an integral part of DNA and protein sequences. In DNA sequences, these internal repeats present in multiple copies are referred as 'direct repeat'. For example, the tandem repeats are repeated consecutively and act as markers for genomic sequence **[1].** In addition, these markers are used for the mapping of genes which are responsible for human genetic diseases **[2].** In eukaryotes, the short DNA tandem repeats present in the telomeric region of chromosome forms the common repeat motif **[3].** Moreover, the internal sequence repeats are known to play an important role in evolution and approximately 14% of the protein sequences contain internal sequence repeats **[4].** Understanding the sequence to structure correlation of internal repeats in protein structures would enhance the knowledge of researchers to further march forward in the world of molecular modeling and *in-silico* drug discovery. Furthermore, eukaryotic proteins are three times more likely to contain internal sequence repeats than prokaryotic proteins **[4]** as these eukaryotic repeats have unique functions. Thus, it is essential to retrieve the information pertaining to internal sequence repeats in an efficient manner which is useful to the research community working in the field of structural biology, molecular modelling and genomics. To this end, we have developed a web server to obtain all the internal sequence repeats in both protein and nucleotide sequences by exploiting the utilities of high-end computing and World Wide Web technologies.

The limitation of the existing programs has already been outlined earlier **[5].** For example, the number of amino acid residues allowed to be entered as input is limited to 1000 in the RADAR web server **[6]** and 2000 in the TRUST web server **[7].** On the other hands, REPPER **[8]** does not pose a limitation on the number of amino acid residues; however, it allows only one protein sequence to be entered at a given time. Further, the recently available web servers SWELFE **[9]** and Censor **[10]** have their own limitations. The SWELFE program permits a single sequence and uses substitution matrix for its search. In the Censor program, repeats scanning are made with the help of reference repeat database, which makes restriction in finding of non annotated internal repeats. To overcome the above issues, FAIR (Finding All Identical Repeats) web server has been deployed to identify the internal sequence repeats in protein as well as DNA sequences. Moreover, the proposed web server FAIR has been tested with a protein sequence of more than 35000 residues available in the genome database. Also, the proposed server accepts many sequences at a time. The advanced features of the proposed web server are discussed in detail in the subsequent sections.

Methodology:
The proposed web server has been implemented with two different algorithms for the identification of identical **[5]** and similar **[11]** repeats. Further, various sequences databases like Genome database, SWISS-PROT **[12]**, UniProt/TrEMBL **[13]**, PIR **[14]** and Protein Data Bank **[15**] are incorporated in the web server. Options are provided to view the three-dimensional structure of the repeats found in the protein data bank. The freely available JAVA plug-in Jmol **[16]** is interfaced for visualizing the three-dimensional structures.

Features:
The proposed computing server requires the sequence of interest and the minimum number of amino acids or nucleotides in a repeat. Further, the users have to specify the input sequence type (Protein or DNA). Also, users can specify the number of occurrences of a repeat. The users need to paste the sequence(s) in FASTA format or upload the sequence(s) from a file. There are options provided to search for either the identical or similar sequence repeats in a protein sequence. Incase of DNA sequences, the above option is limited to identical match only. The server produces a detailed display with a brief summary at the end of the output. The user can save the output produced by the computing server in the hard-disk of their local machine for further analysis.

Implementation:
The computing engine is developed and optimized for Fedora Core (Version 7.0) and is driven by a - 2.66 GHz Quad-core Intel Xeon processor equipped with 4 Gigabytes of Random Access Memory. The operating system is chosen for security and reliability. The computing engine is tested vigorously using all platforms (windows 98/2000, XP, Linux and SGI). As mentioned, for the sequence provided in the case study, the web server displays the results in about two seconds for a single sequence (number of residues = 1436). However, the computation time varies depending upon the network speed. The computing engine is developed using PERL. In addition, Java script is deployed to develop and validate the web pages. For designing the web pages, CSS and HTML have been used. The algorithm used in this computing engine is written in C++ language. The described computing engine is freely available for academic users and non-commercial organizations over the World Wide Web at the URL: http://bioserver1.physics.iisc.ernet.in/fair/. The users are requested to cite this article in their research publications. Please send your comments and suggestions to Dr. K. Sekar (sekar@physics.iisc.ernet.in).

ISSN 0973-2063 (online) 0973-8894 (print)
Bioinformation 4(7): 271-275 (2010)

```
>gi|2896785|emb|CAA17262.1| Probable polyketide synthase pks12 [Mycobacterium tuberculosis H37Rv]
MVDQLQHATEALRKALVQVERLKRTNRALLERSSEPIAIVGMSCRFPGGVDSPEGLWQMVADARDVMSEF....
.....................................................................


Total number of residues present in the input sequence = 4151
---------------------------------------------------------------------
Number of residues in the repeat = 215
APAALRYLSQARHTGKVVMLMPGSWAAGTVLITGGTGMAGSAVARHVVARHGVRNLVLVSRRGPDAPGAA
ELVAELAAAGAQVQVVACDAADRAALAKVIADIPVQHPLSGVIHTAGALDDAVVMSLTPDRVDVVLRSKV
DAAWHLHELTRDLDVSAFVMFSSMAGLVGSSGQANYAAANSFLDALAAHRRAHGLPAISLGWGLWDQASA
MTGGL


The above repeat occurs at [1653 to 1867] and [3683 to 3897]
---------------------------------------------------------------------
Number of residues in the repeat = 128
GQRVLIHAGTGGVGMAAVQLARHLGLEVFATASKGKWDTLRAMGFDDDHISDSRSLEFEDKFRAATGGRG
FDVVLDSLAGEFVDASLRLVAPGGVFLEMGKTDIRDPGVIAQQYPGVRYRAFDLFEPG


The above repeat occurs at [1493 to 1620] and [3523 to 3650]
---------------------------------------------------------------------
Number of residues in the repeat = 307
SEGGGMLVLQRLSDARRLGHPVLAVVVGSAVNQDGASNGLTAPNGPSQQRVVRAALANAGLSAAEVDVVE
GHGTGTTLGDPIEAQALLATYGQDRGEPGEPLWLGSVKSNMGHTQAAAGVAGVIKMVLAMRHELLPATLH
VDVPSPHVDWSAGAVELLTAPRVWPAGARTRRAGVSSFGISGTNAHVIIEAVPVVPRREAGWAGPVVPWV
VSAKSESALRGQAARLAAYVRGDDGLDVADVGWSLAGRSVFEHRAVVVGGDRDRLLAGLDELAGDQLGGS
VVRGTATAAGKTVFVFPGQGSQWLGMG


The above repeat occurs at [267 to 573] and [2290 to 2596]
---------------------------------------------------------------------
```

Figure 1: FAIR server output of identical repeats in the protein sequence from *Mycobacterium tuberculosis H37Rv*.

Case study:

The FAIR web server enables the users to identify the internal sequence repeats in the protein and nucleotide sequences. Following are the various options that recognizes FAIR web server as unique.

Identical repeat in protein sequences

Figure 1 displays a sample output for an input (FASTA format) protein (probable polypeptide synthase pks12) sequence from *Mycobacterium tuberculosis H37Rv*. Identification of repeats in polyketide synthase would be of immense importance as it is an important target for drug designing. The total number of amino acid residues in the input sequence is 4151 (only part of the input is shown in **Figure 1**). The minimum number of residues in the repeat is set to be greater than or equal to 120. The results produced by the computing server are shown in **Figure 1.** The computing engine found three long identical repeats of 215, 128 and 307 length (Figure 1). The results of the proposed server are compared with the SWELFE server for the above input sequence and are reported in **Table 1** (see supplementary material). It is interesting to note that the proposed server identified many identical internal repeats whereas the SWELFE reported only similar repeats (see Table 1 in supplementary material for details).

Identical repeat in nucleotide sequence

The complete genome of *Mycobacterium tuberculosis* H37Rv has been given as the input. The total number of nucleotides in the input sequence is 4411532 (here again, only part of the input sequence is shown in **Figure 2**). For a minimum number of nucleotides in the repeat as 50, the sample output has been displayed in **Figure 2**. These significant large identical repeats would enable scientists to focus on sequence to structure correlation and their implications.

Similar repeats in protein sequence

The given input sequence is taken from the protein Hexokinase 1[*Homo sapiens*], PDB Id: 1CZA **[17]**. The protein consists of 917 amino-acid residues and for a minimum number of residues as ten the output obtained is shown in **Figure 3**. Hexokinase I, a protein found in all mammalian tissues plays an important role as a "house keeping enzyme". Further, insights into the occurrence of such structurally similar repeats would enhance the understanding on the three-dimensional structure of Hexokinase I.

ISSN 0973-2063 (online) 0973-8894 (print)
Bioinformation 4(7): 271-275 (2010)

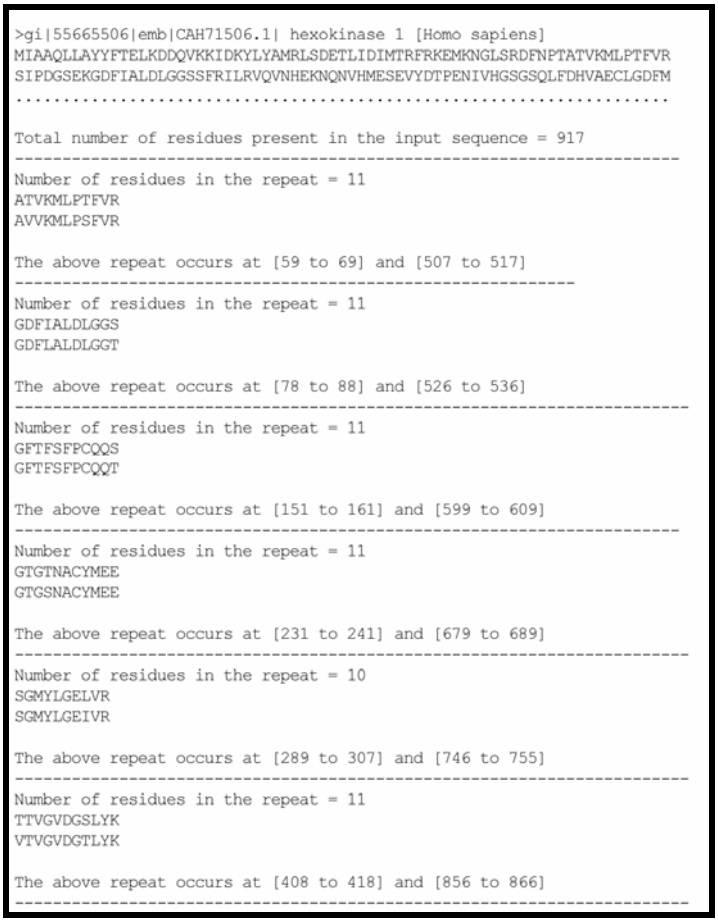
```
>gi|41353971|emb|AL123456.2|MTBH37RV Mycobacterium tuberculosis H37Rv complete genome
TTGACCGATGACCCCGGTTCAGGCTTCACCACAGTGTGGAACGCGGTCGTCTCCGAACTTAACGGCGACC
CTAAGGTTGACGACGGACCCAGCAGTGATGCTAATCTCAGCGCTCCGCTGACCCCTCAGCAAAGGGCTTG
.....................................................................

Total number of residues present in the input sequence = 4411532
---------------------------------------------------------------------
Number of residues in the repeat = 397
CAGGTGGTGGCCTGTGACGCGGCGGATCGAGCGGCGTTGGCCAAGGTGATCGCCGATATTCCGGTTCAGC
ATCCATTGTCGGGCGTGATCCACACCGCCGGCGCACTCGACGACGCGGTGGTGATGTCACTGACACCGGA
TCGGGTGGATGTGGTGTTGCGGTCCAAGGTGGACGCGGCGTGGCACCTGCACGAGTTGACTCGCGACCTG
GATGTGTCGGCGTTTGTCATGTTTTCGTCGATGGCCGGGCTGGTCGGATCGTCGGGCCAGGCCAACTATG
CGGCCGCCAATTCGTTTTTGGATGCGCTGGCCGCCCACCGGCGGGCCCATGGGCTGCCGGCCATCTCCCT
GGGCTGGGGTCTGTGGGATCAGGCCAGCGCCATGACCGGCGGCCTGG

The above repeat occurs at [5206 to 5602] and [11296 to 11692]
---------------------------------------------------------------------
Number of residues in the repeat = 387
CGGGCCAGCGCGTGTTGATCCATGCCGGCACCGGCGGGGTGGGCATGGCGGCGGTGCAGCTGGCTCGGCA
TCTGGGGTTGGAGGTGTTCGCGACCGCGAGCAAGGGTAAGTGGGACACCTTGCGCGCCATGGGCTTTGAC
GACGACCACATATCCGATTCACGTAGCCTAGAGTTCGAGGACAAGTTCCGCGCGGCCACTGGCGGTCGAG
GGTTCGACGTGGTGTTGGACTCGCTGGCCGGTGAATTCGTGGATGCGTCGCTGCGTCTGGTGGCACCGGG
TGGGGTGTTCTTGGAGATGGGCAAGACCGACATCCGCGACCCCGGCGTGATCGCCCAGCAGTACCCGGGC
GTGCGCTACCGCGCCTTCGACCTATTCGAACCGGGAC

The above repeat occurs at [4475 to 4861] and [10565 to 10951]
---------------------------------------------------------------------
Number of residues in the repeat = 66
GAGCTCGCCACGCTGTTCGGCGATGGGGTGCTGCGGCCGTTGCCGGTCACCACTTTTGACGTGCGG
The above repeat occurs at [4888 to 4953] and [10978 to 11043]

---------------------------------------------------------------------
Number of residues in the repeat = 111
GCGATGGATCCACAGCATCGGATGTTGCTGGAGTTGTCCTGGGAGGCGTTGGAGCGGGCCGGTATCGATC
CGACCGGATTGCGCGGCAGCGCCACCGGGGTATTCGCCGGG
The above repeat occurs at [358 to 468] and [6424 to 6534]

---------------------------------------------------------------------
Number of residues in the repeat = 250
GCGCGCCTGCGGCGTTGCGTTATCTGAGCCAGGCGCGCCACACCGGCAAGGTCGTCATGCTGATGCCCGG
CTCGTGGGCGGCGGGCACGGTGTTGATCACCGGTGGCACCGGGATGGCGGGTTCGGCGGTGGCCCGTCAC
GTGGTGGCTCGTCATGGGGTGCGCAATCTGGTGTTGGTGAGCCGGCGCGGCCCGGATGCTCCCGGGGCTG
CGGAGCTGGTGGCCGAGTTGGCCGCCGCCGGTGCGCAGGT
The above repeat occurs at [4955 to 5204] and [11045 to 11294]

---------------------------------------------------------------------
Number of residues in the repeat = 173
TCCAGCGTCGCCTCGGGTCGGGTGGCCTATGTGCTGGGGTTGGAGGGTCCGGCGGTGTCGGTGGATACGG
CGTGTTCGTCGTCGTTGGTGGCGTTGCATATGGCGGTGGGATCGCTGCGGTCGGGGGAGTGCGATCTGGC
GCTGGCTGGCGGCGTCACCGTCAACGCCACACC

The above repeat occurs at [535 to 707] and [6604 to 6776]

---------------------------------------------------------------------
Number of residues in the repeat = 923

TCCGAGGGCGGTGGGATGTTGGTGCTGCAGCGGCTTTCGGATGCGCGGCGGTTGGGTCATCCGGTGTTGG
CGGTGGTGGTCGGGTCGGCGGTTAATCAGGATGGGGCGTCGAATGGGTTGACCGCGCCTAATGGTCCTTC
GCAGCAGCGGGTGGTGCGGGCGGCGTTGGCCAATGCCGGGTTGAGCGCGGCCGAGGTGGATGTGGTGGAG
GGGCATGGGACCGGGACCACGTTGGGGGATCCGATTGAGGCTCAGGCGTTGTTGGCCACTTATGGGCAAG
ATCGGGGGGAGCCGGGAGAACCTTTGTGGTTGGGGTCGGTGAAGTCGAATATGGGTCATACGCAGGCCGC
GGCGGGGGTGGCCGGGGTGATCAAGATGGTGTTGGCGATGCGCCATGAGCTGTTGCCGGCGACGTTGCAC
GTGGATGTGCCTAGCCCGCATGTGGATTGGTCGGCGGGGGCGGTGGAGTTGTTGACCGCGCCGCGGGTGT
GGCCTGCTGGTGCTCGGACGCGTCGTGCGGGGGTGTCGTCGTTTGGGATTAGTGGCACTAATGCGCATGT
GATTATCGAGGCGGTGCCGGTGGTGCCGCGGCGGGAGGCTGGTTGGGCGGGGCCGGTGGTGCCGTGGGTG
GTGTCGGCGAAGTCGGAGTCGGCGTTGCGGGGGCAGGCGGCTCGGTTGGCCGCGTACGTGCGTGGCGATG
ATGGCCTCGATGTTGCCGATGTGGGGTGGTCGTTGGCGGGTCGTTCGGTTTTTGAGCATCGGGCGGTGGT
GGTTGGCGGGGACCGTGATCGGTTGTTGGCCGGGCTCGATGAGCTGGCGGGTGACCAGTTGGGCGGCTCG
GTTGTTCGGGGCACGGCGACTGCGGCGGGTAAGACGGTGTTCGTCTTCCCCGGCCAAGGCTCCCAATGGC
TGGGCATGGGAAT


The above repeat occurs at [799 to 1721] and [6868 to 7790]
---------------------------------------------------------------------
Number of residues in the repeat = 59

TGGTGCGGTCGGCGCAGACCGAGCATCCGGGCCGGATCGTGCTGGTCGATTCCGATGCG

The above repeat occurs at [3872 to 3930] and [9959 to 10017]
---------------------------------------------------------------------
```

Figure 2: Different lengths of identical repeats identified in the complete genome of Mycobacterium tuberculosis H37Rv.

Conclusion:

The proposed computing server displays all identical and similar repeats in the provided input sequence. There is no limitation on the number of residues in the input sequence. The interesting feature is that the user will be able to find the occurrence of the resultant repeat in other sequence and structure databases, if any such repeat is present. Further, the three-dimensional structure of the corresponding repeat (identical or similar) can be visualized in the local machine. The computing engine will be an interactive tool to those working in the area of structural biology, molecular modeling and Bioinformatics.

ISSN 0973-2063 (online) 0973-8894 (print)
Bioinformation 4(7): 271-275 (2010)



```
>gi|55665506|emb|CAH71506.1| hexokinase 1 [Homo sapiens]
MIAAQLLAYYFTELKDDQVKKIDKYLYAMRLSDETLIDIMTRFRKEMKNGLSRDFNPTATVKMLPTFVR
SIPDGSEKGDFIALDLGGSSFRILRVQVNHEKNQNVHMESEVYDTPENIVHGSGSQLFDHVAECLGDFM
.....................................................................

Total number of residues present in the input sequence = 917
---------------------------------------------------------------------
Number of residues in the repeat = 11
ATVKMLPTFVR
AVVKMLPSFVR

The above repeat occurs at [59 to 69] and [507 to 517]
-------------------------------------------------------------
Number of residues in the repeat = 11
GDFIALDLGGS
GDFLALDLGGT

The above repeat occurs at [78 to 88] and [526 to 536]
---------------------------------------------------------------------
Number of residues in the repeat = 11
GFTFSFPCQQS
GFTFSFPCQQT

The above repeat occurs at [151 to 161] and [599 to 609]
---------------------------------------------------------------------
Number of residues in the repeat = 11
GTGTNACYMEE
GTGSNACYMEE

The above repeat occurs at [231 to 241] and [679 to 689]
---------------------------------------------------------------------
Number of residues in the repeat = 10
SGMYLGELVR
SGMYLGEIVR

The above repeat occurs at [289 to 307] and [746 to 755]
---------------------------------------------------------------------
Number of residues in the repeat = 11
TTVGVDGSLYK
VTVGVDGTLYK

The above repeat occurs at [408 to 418] and [856 to 866]
---------------------------------------------------------------------
```

Figure 3: Similar sequence repeats obtained in the protein Hexokinase1 (1CZA).

Acknowledgements:
All the authors acknowledge the use of the Bioinformatics Centre, the Interactive graphics facility and the Supercomputer Education and Research Centre. One of the authors (KS) thanks the Department of Information Technology (DIT) for financial support. The authors (NB and NC) thank Dr. K. Sekar for providing an opportunity to undergo internship program during the summer of 2006.

References:

[1] A. M. Klevytska *et al., J. clinical Microbiology*, 39: 3179 (2001) [PMID: 11526147]
[2] Y. Nakamura *et al., Science*, 235: 1616 (1987)
[3] W. Traut *et al., Chromosome Research*, 15: 371 (2007) [PMID: 17385051]
[4] E. M. Marcotte *et al., J. Mol. Biol.*, 293: 151 (1999) [PMID: 10512723]
[5] N. Banerjee *et al., Curr. Sci.* 95: 188 (2008)
[6] A. Heger. & L. Holm, *Proteins*, 41: 224 (2000) [PMID: 10966575]
[7] R. Szklarczyk & J. Heringa, *Bioinformatics*, 20:311 (2004) [PMID: 15262814]
[8] M. Gruber *et al., Nucl. Acids. Res.*, 33: 239 (2005) [PMID: 15980460]
[9] L. Abraham *et al., Bioinformatics*, 24: 1536 (2008) [PMID: 18487242]
[10] O. Kohany *et al., BMC Bioinformatics*, 7: 474 (2006) [PMID: 17064419]
[11] N. Banerjee *et al., Curr. Sci.* 97, 1345 (2009).
[12] A. Bairoch & R. Apweiler, *Nucl. Acids. Res*., 26: 38 (1998) [PMID: 10592178]
[13] R. Apweiler *et al., Nucl. Acids. Res*. 32: 115 (2004) [PMID: 14681372]
[14] W. C. Barker *et al., Nucl. Acids. Res*., 26: 27 (1998)
[15] H. M. Berman *et al., Nucl. Acids. Res*., 28: 235 (2000)
[16] http://www.jmol.org
[17] E. Aleshin *et al., J. Mol. Biol*. 296: 1001 (2000) [PMID: 10686099]

Edited by P. Kangueane
Citation: Senthilkumar *et al.*, Bioinformation 4(7): 271-275 (2010)

ISSN 0973-2063 (online) 0973-8894 (print)
Bioinformation 4(7): 271-275 (2010)

Supplementary material:

Table 1: Comparison of internal repeats produced by FAIR and SWELFE web servers for a protein sequence (probable polypeptide synthase pks12 from *Mycobacterium tuberculosis H37Rv*).

Repeats identified by FAIR	Number of amino-acids in the repeat	Repeat identified by SWELFE	Number of amino-acids in the repeat
115 – 156	42	1238 – 1276	39
2137 – 2178		2950 – 2989	
179 – 236	58	2659 – 2759	109
2202 – 2259		3705 – 3813	
1493 – 1620	128	1675 – 1783	109
3523 – 3650		2659 – 2759	
1653 – 1867	215	-	-
3683 – 3897			
267 – 573	307	-	-
2290 – 2596			

ISSN 0973-2063 (online) 0973-8894 (print)
Bioinformation 4(7): 271-275 (2010)