**International Journal of Advanced Research in Engineering and Technology (IJARET)**
Volume 9, Issue 4, July - August 2018, pp. 1–5, Article ID: IJARET_09_04_001
Available online at http://www.iaeme.com/IJARET/issues.asp?JType=IJARET&VType=9&IType=4
ISSN Print: 0976-6480 and ISSN Online: 0976-6499
© **IAEME** Publication

# MOTIF FINDING TOOL USING PERL CGI

**Arundhati Mahesh**
Faculty of Biomedical Sciences, Technology & Research,
Sri Ramachandra Medical College and Research Institute, Porur, Chennai, Tamilnadu, India

**G Dicky John Davis**
Faculty of Biomedical Sciences, Technology & Research,
Sri Ramachandra Medical College and Research Institute, Porur, Chennai, Tamilnadu, India

**PK Ragunath**
Faculty of Biomedical Sciences, Technology & Research,
Sri Ramachandra Medical College and Research Institute, Porur, Chennai, Tamilnadu, India

**E Kannan**
VelTech Rangarajan Dr. Saguthala R&D Institute of Science and Technology,
Avadi, Chennai, Tamilnadu, India

## ABSTRACT

*Bioinformatics is an assimilation between biology and information technology. Computers have always played an important role in biology since their invention and development over the last few epochs. They are now extensively used to store, process and modulate biological data. Several programs have been written to develop tools and applications for gene-finding, to identify restriction sites, perform restriction digests, create restriction maps, identify exons and introns, and also create models and pathways of cellular processes. Motif finding is one of the common undertakings in bioinformatics and several motif finding programs have been formulated. A gene is a functional unit of the human genome and is found in DNA. DNA is transcribed to give mRNA which is translated to produce proteins. Proteins are made up of amino acids coded for by the mRNA. Hence, the production of proteins depends in the specific set of DNA sequences that are a part of a gene. DNA has several pattern specific and profile specific features, one of which includes motifs. A motif is a short DNA or protein sequence that contributes to the biological function of the sequence in which it resides. Thus the most fundamental studied problem is the discovery of motifs in sequences computationally. Thus, the motif finding tool was developed using Perl CGI and word based algorithms which exhaustively searches through the input sequences and protein sequences. The developed tool does not require any trained data set like the other motif finding tools. This motif finding tool has been designed to find sequence specific motifs and also generate size specific motifs with positions in nucleotide and protein sequences.*

**Key words:** Bioinformatics, Motif finding, DNA, PERL CGI, Algorithm.

**Cite this Article:** Arundhati Mahesh, G Dicky John Davis, PK Ragunath and E Kannan, Motif Finding Tool using PERL CGI. *International Journal of Advanced Research in Engineering and Technology,* 9(4), 2018, pp 1–5.
http://www.iaeme.com/IJARET/issues.asp?JType=IJARET&VType=9&IType=4

## 1. INTRODUCTION

Information technology has always played an important role in biology since their invention and development over the last few decades. Motifs are recurring patterns found in DNA and protein sequences. DNA and protein sequence motifs are presumed to have biological significance and they can be used to predict protein function. Protein sequences also contain structural motifs that may be conserved in a large number of different proteins, contributing towards the formation of their three dimensional structures. Sequence motifs can be used to indicate nucleases and transcription factors by sequence specific binding sites for proteins and also involves study in proteins with important developments at the RNA level, ribosome binding, mRNA processing including splicing, editing, polyadenylation and transcription termination (D'haeseleer & Patrik 2006). Motifs in a DNA or a protein sequence are not always specific, but they can be present in several pattern-specific or position specific variations. Biological motifs can comprise of short DNA motifs, longer protein motifs and evolutionary significant recurring motifs. The result of convergent evolution can be due to primary comprises of short motifs often found at functional sites of biopolymers mainly cleavage sites, binding sites and attachment sites. The arising of divergent evolution can be due to the second comprises longer protein motifs associated with globular structural domains. The final recurring motifs can rise from evolutionarily recent duplications, such as DNA transposons(Frith et al. 2008). These are usually short in length and have a high level of sequence variability and cannot be reliably predicted by computational means. Therefore, when there is an experimental evidence that the motif is functionally important or its presence is constant with the purpose of the protein, only then putative motifs are noted. An incomplete understanding of the biology of regulatory mechanism does not always provide adequate evaluation of underlying algorithms over motif models.

Over past years, many computational methods have been defined for identifying, characterizing and searching with sequence motifs. There are a variety of motif finding tools and software available that have different approaches to motif finding.

**MEME Suite:** A unified web server interface used in search of similar motifs in databases of known motifs and associations between motifs is developed as a software toolkit using ANSI C as command line interface and later published as SOAP web services.

**oPOSSUM:** An integrated tools for analysis of regulatory motif over-representation which is an internet-based system.

**FANMOD:** The tool used in motif detection in colored networks which got enhanced with the improvement in efficiency of network motif detection by some orders of magnitude on an existing tool with more reliability in fast network motif detection.

**FIMO:** The program calculates a log-likelihood ratio score for each position in a given sequence database and thereby converts score to P-value programmatically and then applies false discovered rate analysis to estimate a Q-value. Thus DNA scanning or protein sequences with motifs is described as position specific scoring matrices in the FIMO software.

Motif finding programs are designed to find motifs depending on their pattern or profile based on the sequence. Using bioinformatics components to conduct motif search has proved to be very time efficient and an easily executable process. Manual motif finding can be incredibly time consuming and almost impossible for large sequences. Computer programs do the same task quickly, without errors. Because Perl has several features that help find patterns and letters in strings, it is a popular language used when it comes to motif finding. Perl programs read the input sequences, calculate its length, pattern search the string letter wise or number wise and provide a desired output. Most motif finding programs require two inputs, the query sequence and the motif. The motif input can be in the form of letters or numbers. The programs that require oligonucleotides as an input, calculates the length of the oligonucleotide, finds the pattern in the query sequence, calculate how many times the pattern is repeated and provides the result. Numerical input programs identify the motif length based on the input and search the query sequence for repeated patterns of that size and provide the result.

The motif finding tool was designed using the word-based algorithm that exhaustively finds desired motifs in the input nucleotides and protein sequences. Parameters like specific motifs, sequence length and experimentally found motifs are considered in this tool.

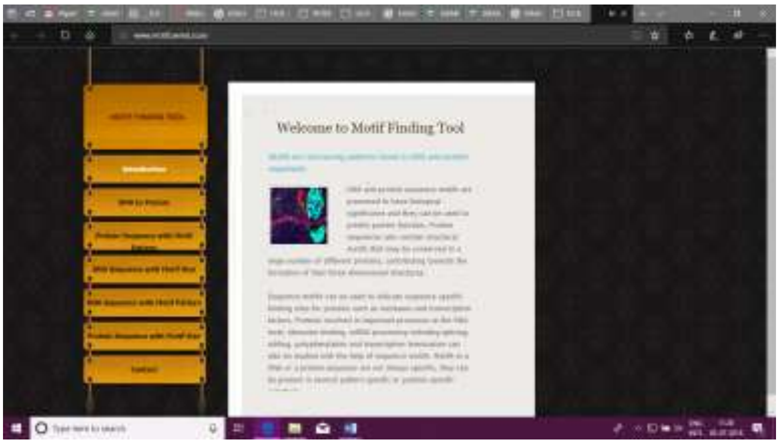
## 2. METHODOLOGY

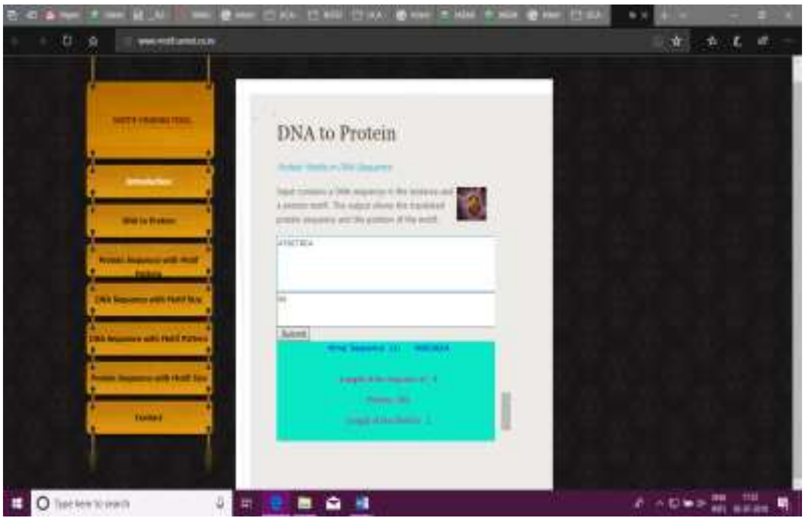
The Motif finding tool is developed using Perl CGI by configuring it in WAMP server and there by developing a front end using Java scripting and the results are displayed in a understandable as well as tabular format using HTML and CSS. This tool can be accessed at http://www.motif.swmd.co.in/

## 3. RESULTS

This tool offers a pattern finding search based on the input sequence and a profile based search to find the position of repeated motifs based on the input size. It finds the motif pattern or size in the input nucleotide or protein sequence. This could be advantageous to users that have a single input sequence and know the size or the pattern they need to search in the sequence. This tool does not require extracting sequences from a database and no parameters have to be set in this tool, unlike several other motif finding tools available which makes a quick and easy way to find the desired motif. This tool can be used to find several known motifs in nucleotide sequence of any organism. These motifs can be promoter sequences, palindromic sequences, etc. One application of this would be searching the TATA box promoter sequence in bacterial genomes.

The input would require a bacterial nucleotide sequence and the TATA box motif sequence. The output displays the position of the motif in the nucleotide sequence. Protein specific motifs like the DEAH box, found in helicase coding genes can be found in the protein sequence. Motifs can be found based on their size using this tool. If the pattern of the motif is unknown but the size of the motif is known, one can find the motif in the sequence using the tool. This can be useful to find similar sequence motifs that are of the same size. For example, the DEAH box belongs to the family of DEXX motifs which contains similar four letter motifs like DEAD, DECH, DEYQ, etc. (www.uniprot.com).





## 4. DISCUSSION

Motifs are recurring patterns found in nucleotide and protein sequences. These motif sequences are presumed to have biological significance in relation to sequences and structures. The motif finding tool was designed using Perl programming language. It searches the input sequence exhaustively using word based algorithms. Sequence motifs can be found in nucleotide and protein sequences based on the sequence and the size of the motif. It can also be used to find a protein specific motif in a nucleotide sequence. The motif finding tool does not require a trained data set like the other motif finding tools that often use heuristic algorithms and Hidden Markov Models for motif search. Parameters like specific motifs, sequence length and experimentally found motifs are considered in this tool.

## 5. CONCLUSIONS

Thus, this tool can be used to find sequence specific and size specific motifs in a wide range of nucleotide and protein sequences. It can also be used to find a protein specific motif in a nucleotide sequence. This motif finding tool can be useful while detecting motifs in nucleotide and protein sequences in biological studies.

## REFERENCES

[1] Alberts, B. et al., 2014. Molecular Biology of the Cell, Garland Publishing.

[2] Alon, U. & Uri, A., 2007. Network motifs: theory and experimental approaches. Nature reviews. Genetics, 8(6), pp.450–461.

[3] Anon, Available at: http://biochem218.stanford.edu/Projects%202012/Lin.pdf [Accessed April 5, 2016a].

[4] Anon, Website. Available at: http://www.plosone.org/article/fetchSingleRepresentation.action?uri=info:doi/10.1371/journal.pcbi.1000832.s001. [Accessed April 5, 2016b].

[5] Bailey, T.L. et al., 2006. MEME: discovering and analyzing DNA and protein sequence motifs. Nucleic acids research, 34(Web Server), pp.W369–W373.

[6] Ben-Hur, A., Asa, B.-H. & Douglas, B., Sequence Motifs: Highly Predictive Features of Protein Function. In Studies in Fuzziness and Soft Computing. pp. 625–645.

[7] Bernard, V. et al., 2010. TC-motifs at the TATA-box expected position in plant genes: a novel class of motifs involved in the transcription regulation. BMC genomics, 11(1), p.166.

[8] Das, M.K. & Ho-Kwok, D., 2007. A survey of DNA motif finding algorithms. BMC bioinformatics, 8(Suppl 7), p.S21.

[9] D’haeseleer, P. & Patrik, D. ’haeseleer, 2006. What are DNA sequence motifs? Nature biotechnology, 24(4), pp.423–425.

[10] Frith, M.C. et al., 2008. Discovering Sequence Motifs with Arbitrary Insertions and Deletions. PLoS computational biology, 4(5), p.e1000071.

[11] Nucleic Acids Research, 2007, Vol. 35, Web Server issue W245–W252 doi:10.1093/nar/gkm427