Journal of Computational Intelligence in Bioinformatics
ISSN: 0973-385X Volume 4 Number 1 (2011) pp. 83-93

http://www.ripublication.com/jcib.htm

MSAP: Multiple Sequence Alignment Processor with Block and Pattern Analyzer.

Krishna Patel and Hetalkumar J. Panchal*

**G.H. Patel Post Graduate Department of Computer Science & Technology,*
Sardar Patel University, Vallabh Vidyanagar-388120,
Dist: Anand, Gujarat, India

Abstract

Multiple Sequence Alignment Processor (MSAP) with Block and Pattern Analyzer consist of three modules. MSAP tool which extracts complete identical blocks from output file generated by Multiple Sequence Alignment tool CLUSTALW with the extension .aln. Block finder will find the block from the query sequence and Pattern scan extracts desired pattern from the query sequence. The output file of MSAP and Block Finder tool contain extracted identical blocks, their length, relative position and absolute position in the query sequence. The output file of Pattern scan contains extracted matching pattern, their length, relative position and absolute position in the query sequence. Block Finder and Pattern Scan focus on finding short conserved region in sequences which may represent conserved structure and biological functions.

Keywords: Block, Pattern, ClustalW, Conserved region.

Introduction

The availability of biological sequence data is one of the major consequences of the revolution brought by high throughput biology. Large scale DNA sequencing projects now routinely produce huge amounts of DNA sequences, and the protein sequences deduced from them. Sequence blocks and patterns in biological sequences can have functional or structural implications such as regulatory regions, transcription sites, family signatures, etc. This assumption has been used to search for exceptional words in a particular genome. (Gregory Nuel et al.)[1]. When a protein's function cannot be experimentally determined, it can often be inferred from sequence similarity. Should this process fail, analysis of the protein structure can provide functional clues or confirm tentative functional assignments inferred from the sequence.(James D Watson et al.)[2]

The Clustal series of programs are widely used in molecular biology for the multiple alignment of both nucleic acid and protein sequences and for preparing phylogenetic trees. The popularity of the programs depends on a number of factors, including not only the accuracy of the results, but also the robustness, portability and user-friendliness of the programs. New features include NEXUS and FASTA format output, printing range numbers and faster tree calculation. (Chenna R. et al.)[3]

As per our knowledge, available tools processes the alignment file of CLUSTALW and extracts all the ungapped conserved sequences from alignment file designated as BLOCKS based on minimum length and maximum length provided by user. The putative blocks so obtained are cross validated with interpro (Database of annotated blocks) for their significance and result is obtained in block format. But it does not extract the complete identical blocks.

In this paper, we describe MSAP (Multiple Sequence Alignment Processor), a block extraction algorithm and pattern finding algorithm to extract block from CLUSTALW output file with extension .aln and to find the pattern from the query sequence. MSAP with Block and Pattern Analyzer allows user to extract complete identical blocks of length ranging from 8 to 99 from the output file with extension .aln generated by CLUSTALW. Block Finder aids to search the extracted blocks in query sequence or sequences of fasta format. Pattern scan can also be carried out on one or more sequences for more inference about presence of desired pattern.

Program Feature Summary

- MSAP with Block and Pattern Analyzer detects the complete identical blocks in the output file with the extension .aln generated by CLUSTALW and also aids to search the extracted blocks in other query sequences present of fasta format.
- MSAP with Block and Pattern Analyzer include three utilities:
    - MSAP
    - Block Finder
    - Pattern Scan

- Length of block depends upon the length provided by user.
- The output file contains information about number of query sequences analyzed, number of blocks or matching pattern, and their length, absolute position and relative position in the sequences.
- If user wishes to generate a permanent output file with .doc extension then user has to provide a filename at the homepage which contains detail result output. These files contain information like input parameters, number of sequence analyzed, number of blocks or matching pattern, and their length, absolute position and relative position in the sequence.

Materials and Methods

Requirements

Completely installed ACTIVEPERL 5.0 or updated version. Perl is free package and is licensed under the Artistic License and the GNU General Public License. Distributions are available for most operating systems. Users of Microsoft Windows typically install one of the native binary distributions of Perl for Win32, most commonly Strawberry Perl or ActivePerl. PERL for Macintosh is named as MacPerl. It is supported on many types of Unix except PDP-11s.

Input Prerequisites for Tool

MSAP accepts output file of CLUSTALW with extension .aln saved in text file.
Block Finder and Pattern Scan accept the sequence or sequences in fasta format saved in text file.

Methodology

CGI (Common Gateway Interface) scripting is used with PERL (Practical Extraction and Reporting Language) to build up this tool. PERL provides the powerful text processing facilities without the arbitrary data length limits. Perl has many other features that ease the task at the expense of greater CPU and memory requirements. Some of these are as follows: (Steven Holzner)[4]

- Automatic memory management
- Dynamic typing
- Strings
- Arrays
- Regular expressions

The Common Gateway Interface (CGI) defines how web server software can delegate the generation of web pages to a console application. Such applications are known as CGI Scripts. In simple words the CGI provides an interface between the web servers and the clients. They will identify the request from client and will invoke appropriate function to return the result to the clients. (Ivan Bayross)[5]

Algorithms and Results

The MSAP block extraction and pattern searching algorithm uses different approach from existing tools, which we believe has great potential for block and pattern extraction from the query sequence.

Overview of MSAP Algorithm

First it extracts the aligned sequences from the alignment file of CLUSTALW using regular expression. Extracted sequences are then stored in 2-D array. Every sequence is scanned keeping the row position fixed of the first sequence of alignment file, and search for identical character in respective column. If identical characters are not found then appropriate error message will be displayed. If the identical character

found will be then it will be pushed in other 1-D array else dot will be pushed. The array variable so obtained will be converted to scalar variable. The scalar variable will be filtered with all dots inserted using regular expression, henceforth we will get all the aligned blocks. This aligned block will then be filtered on basis of length entered by user. Length if the block will be measure using in-built function. Relative position will be identified using regular expression and absolute position will be identified using mathematical sequence. And the result will be displayed. (fig.4)

Overview of Block Finder

First, it extracts the sequence/sequences from fasta format and stores it in a scalar variable. User entered block will be searched in sequence using regular expression. If blocks are not found then appropriate error message will be displayed else length of the block will be measure using in-built function. Relative position will be identified using regular expression and absolute position will be identified using mathematical sequence and the result will be displayed. (fig.7)

Overview of Pattern Scan

First, it extracts the sequence/sequences from fasta format and stores it in a scalar variable. User entered pattern will be searched in sequence using regular expression. If matching pattern is not found then appropriate error will be displayed else length of the pattern matching segment will be measure using in-built function. Relative position will be identified using regular expression and absolute position will be identified using mathematical sequence and the result will be displayed. (fig.10)

MSAP with Block and Pattern Analyzer

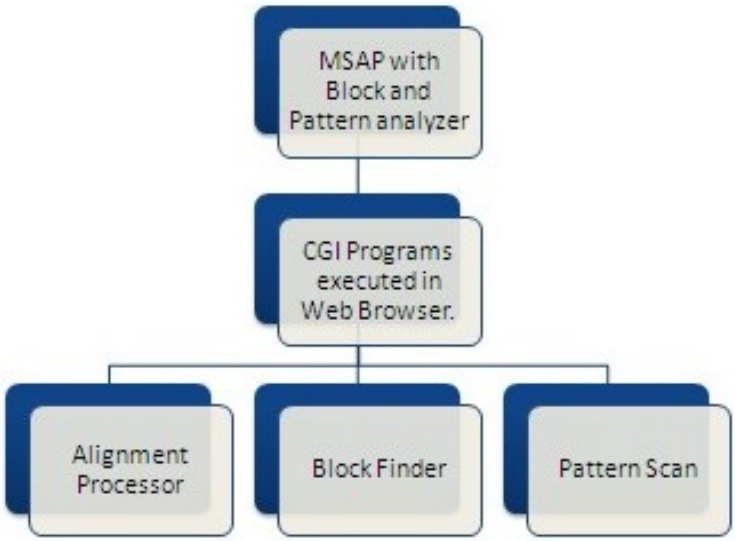

Figure 1: Modules of MSAP with Block and Pattern Analyzer.

Detailed Flowchart of CGI Execution of MSAP with Block and Pattern Analyzer

Figure 2: Flowchart of CGI script execution.

Homepage of MSAP with Block and Pattern Analyzer

Different modules of MSAP with Block finder and Pattern Scan are accessible from icons provided in the left frame of the homepage. (fig.3)

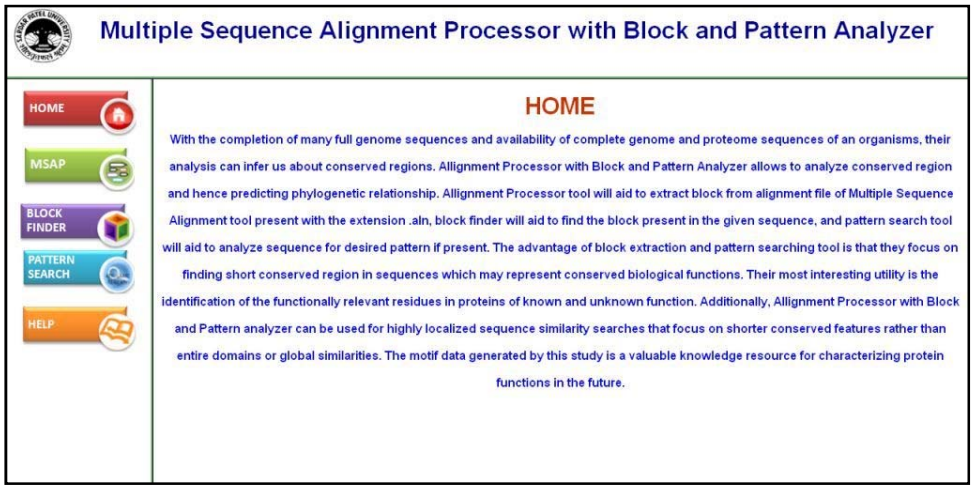

Figure 3: Homepage of MSAP with Block and Pattern Analyzer.

Detailed Flowchart of MSAP CGI Script Execution

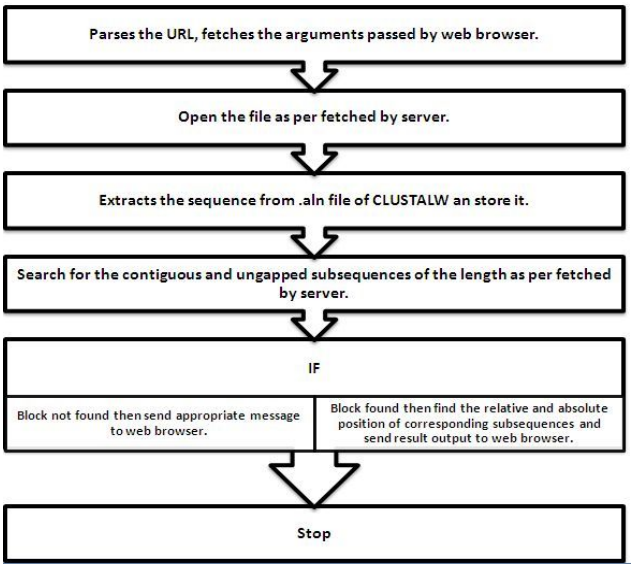

Figure 4: Flowchart of MSAP CGI script execution. This execution takes place on server.

MSAP

To use the proposed tool, one has to generate .aln file by using CLUSTALW (alignment sequence alignment). Generated output file should be saved in text file. Only one .aln file should be saved in single text file. More than one .aln file in single text file would result in invalid file format. This text file should be submitted in the space provided on the MSAP window. To process the input, click upload button (fig.5)

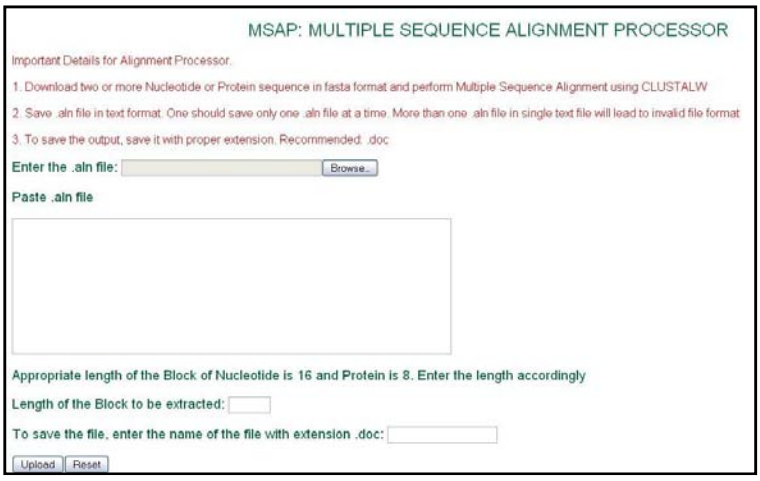

Figure 5: MSAP Input window.

Output of MSAP

Output would be generated on same screen with the detail of number of sequence analyzed, input parameters, extracted blocks, their length, absolute position and relative position. (fig.6)

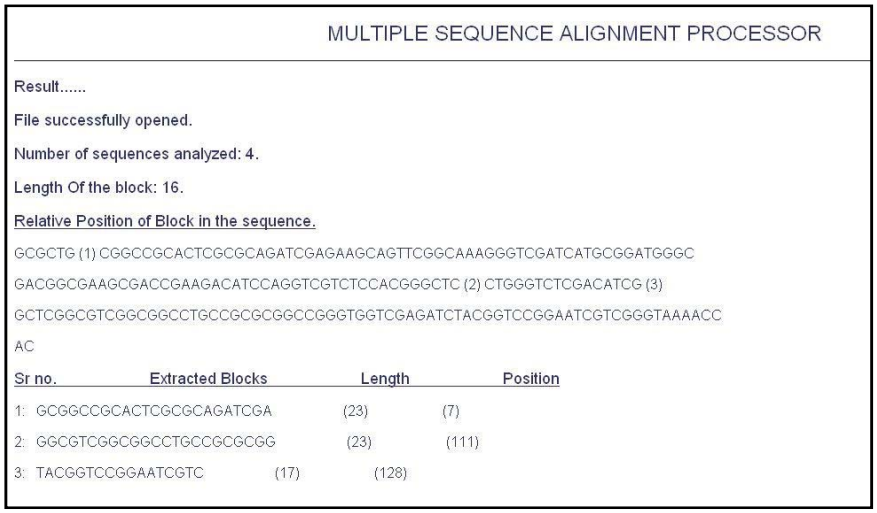

Figure 6: MSAP output window.

Detailed Flowchart of MSAP Block Finder CGI Script Execution

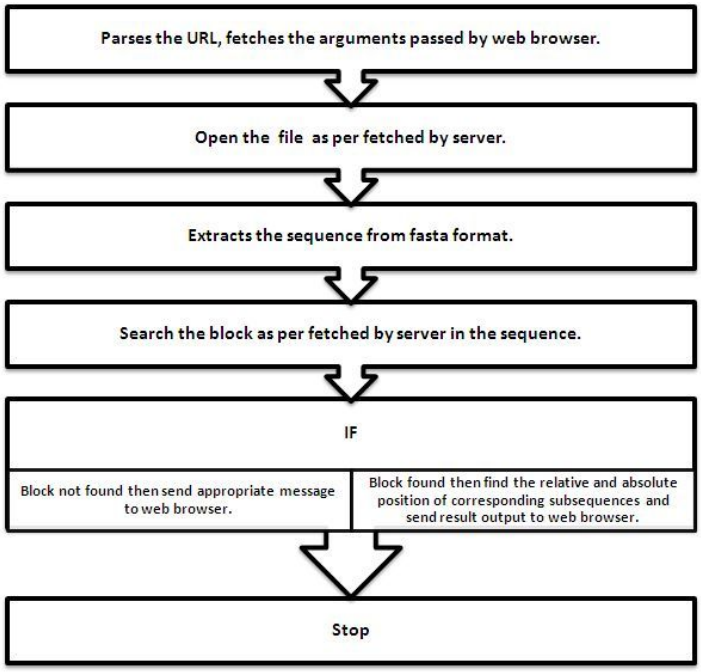

Figure 7: Flowchart of Block Finder CGI script execution.

MSAP Block Finder

To use the proposed tool, one has to fetch sequences in fasta format. More than one senquence of fasta format can be saved in single text file. This text file should be submitted in the space provided on the Block finder window. To process the input, click upload button (fig.8)

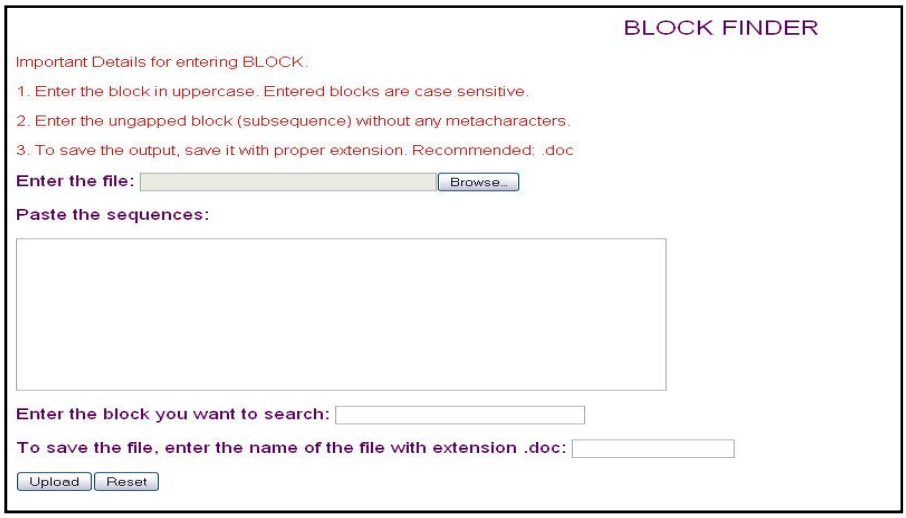

Figure 8: MSAP Block Finder input window.

Output of MSAP Block Finder

Output would be generated on same screen with the detail of number of sequence analyzed, input parameters, extracted blocks, their length, absolute position and relative position. (fig.9)

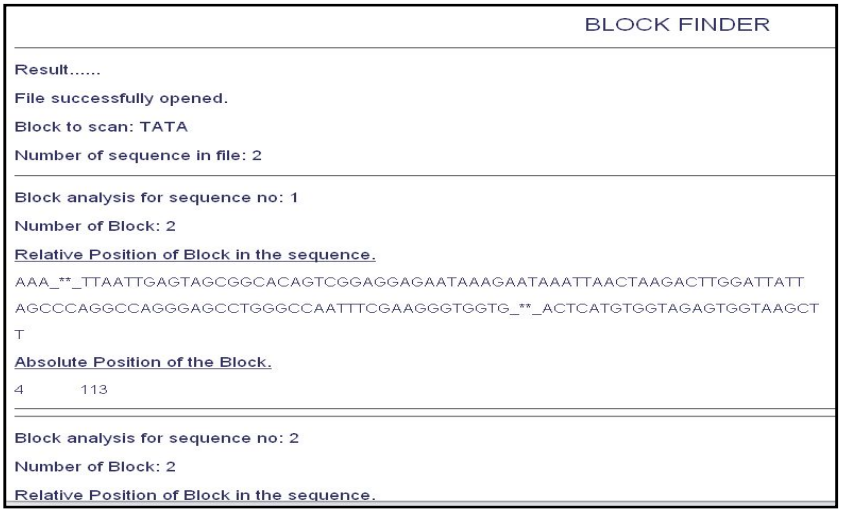

Figure 9: MSAP Block Finder output window.

Detailed Flowchart of MSAP Pattern Scan CGI Script Execution

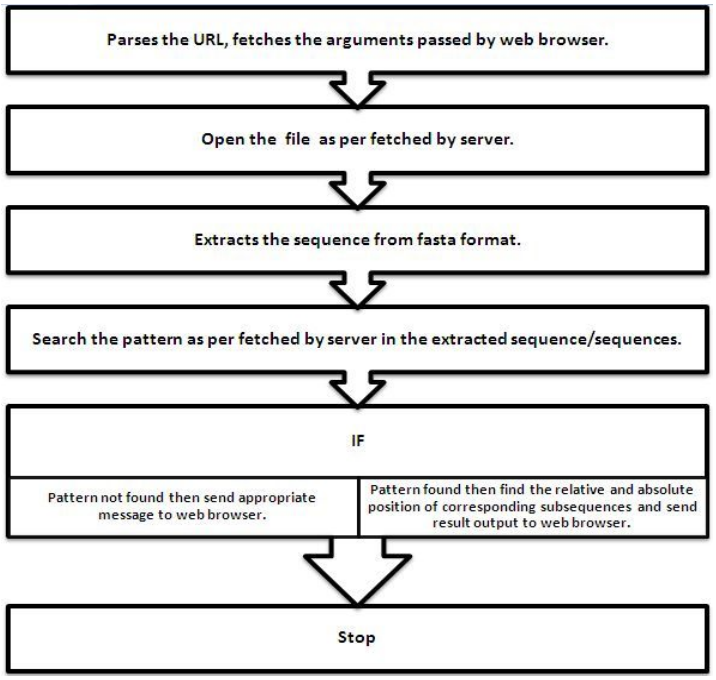

Figure 10: Flowchart of MSAP Pattern Scan CGI script execution.

MSAP Pattern Scan

To use the proposed tool, one has to fetch sequences in fasta format. More than one senquence of fasta format can be saved in single text file. This text file should be submitted in the space provided on the Block finder window. To process the input, click upload button (fig.11)

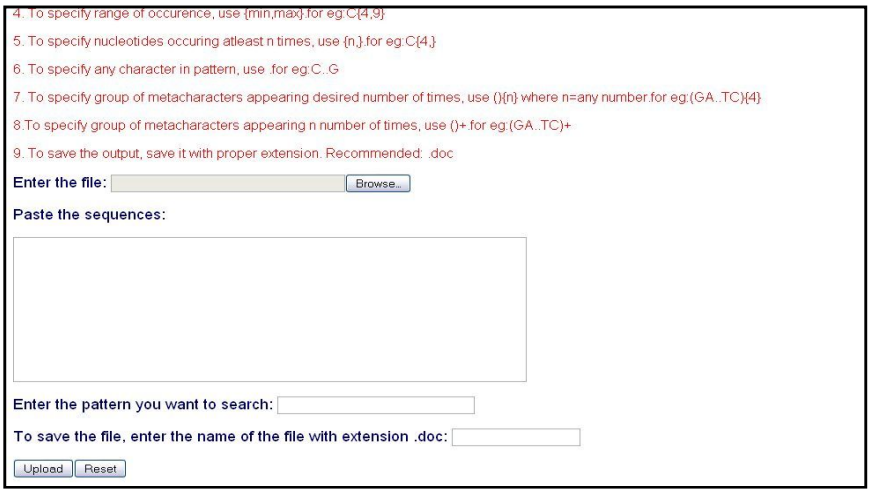

Figure 11: MSAP Pattern Scan input window.

Output of MSAP Pattern Scan

Output would be generated on same screen with the detail of number of sequence analyzed, input parameters, extracted blocks, their length, absolute position and relative position. (fig.12)

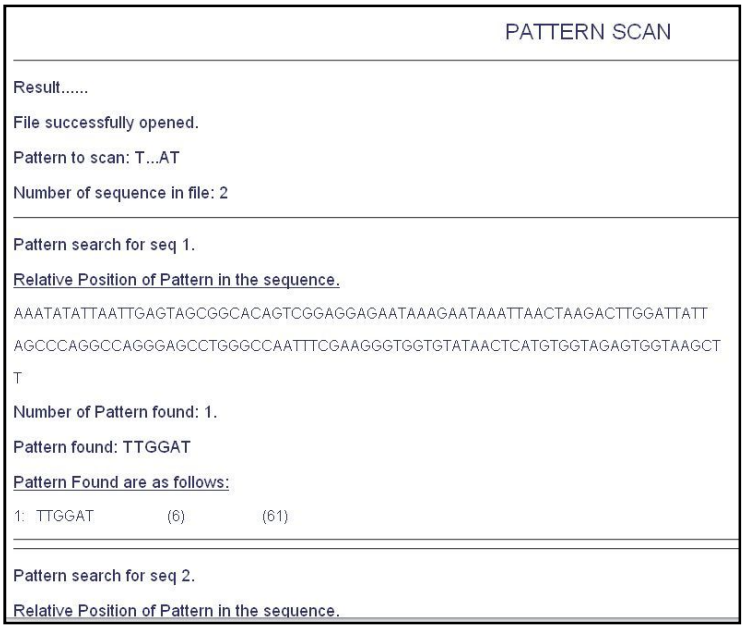

Figure 12: MSAP Pattern Scan output window.

Discussion

Block is ungapped sequence with each row a different DNA or protein segment and each column an aligned nucleotide or residue position. Approach used to extract block is completely different from that used in Block Multiple Sequence Alignment Processor. Identical blocks so obtain from alignment file represent the segment shared by sequences. There are unidentical sequence blocks present in the alignment file which are not extracted by MSAP with Block and Pattern Analyzer. Identical protein sequence will lead to identical structure and function. Unidentical protein blocks having amino acid of same class, can also have similar structure and function. Generally distinctly different protein sequences are taken into consideration when analyzing unidentical blocks. There are certain important words, which remain conserved throughout evolution. There is chance of substitution in such highly conserved block with the amino acid substitution of same class. So they lead to similar protein structure, hence there is no alteration in the function of the protein. Identical blocks lead us to unambiguous result about the structure and function of protein sequence as it has not undergone any substitution. Hence if we come across the complete identical block of sequence in alignment of related or unrelated sequences, having some biological function, we can conclude that they must be having same protein structure. Identification of this identical region will help in

annotating the protein and helps in converting database to knowledgebase. MSAP with Block Finder and Pattern Analyzer can also be useful in clustering the protein on the basis of identical region present in the sequence.

Conclusion

There are various tools available for performing alignments and getting the inferences, but there is no existing tool which allows you to process the alignment file for extracting the complete identical block. There are applications which allow user to extract conserved block from alignment file but not for identical block.

Unidentical blocks so extracted may or may not be of significant use, but identical blocks acknowledge us about sequences similarity, hence the sharing of the sequences, and henceforth the phylogenetic relationship. As mentioned similar sequences are also responsible for similar structure and function. Thus the present tool is very useful in the bioinformatics analysis of sequences.

Acknowledgment

The authors thank Dr. D. B. Choksi Professor and Director (G H Patel Post Graduate Department of Computer Science & Technology, Sardar Patel University, Vallabh Vidyanagar, Gujarat.) for providing required facilities and support to do the present work.

References

[1] Gregory Nuel, Leslie Regad, Juliette Martin, Anne-Claude Camproux: Exact distribution of a pattern in a set of random sequences generated by a Markov source: applications to biological data, *Algorithms for Molecular Biology*, 5:15-33

[2] James D Watson, Roman A Laskowski, Janet M Thornton: Predicting protein function from sequence and structural data. *Current Opinion in Structural Biology*, 15:275-284

[3] Chenna R, Sugawara H, Koike T, Lopez R, Gibson TJ, Higgins DG, Thompson JD: "Multiple sequence alignment with the Clustal series of programs". *Nucleic Acids Res*, 31 (13): 3497–3500

[4] Steven Holzner: PERL BLACK BOOK.

[5] Ivan Bayross: HTML, DHTML, PERL