African Journal of Biotechnology Vol. 11(34), pp. 8414-8416, 26 April, 2012
Available online at http://www.academicjournals.org/AJB
DOI: 10.5897/AJB12.427
ISSN 1684–5315 © 2012 Academic Journals

Review

BioMatriX: Sequence analysis, structure visualization, phylogenetics and linkage analysis workbench

Shagufta Kanwal[1]*, Usman Ali[2], Muhammad Irfan Khan[1], Zainab noor[1], Farhat-ul-ain Mirza[1]

[1]International Islamic University, H-10, Islamabad, Pakistan.
[2]Lahore University of Management Sciences, DHA, Lahore Cantt, Pakistan.

Accepted 14 March, 2012

The BioMatriX (Build Mine Xplore) is a bioinformatics work bench (http://www.bmx-biomatrix.blogspot.com) developed for biological science community to augment scientific research regarding genomics, proteomics, phylogenetics and linkage analysis in one platform. BioMatriX offers multi-functional services to perform specific tasks like DNA/RNA/Protein sequence analysis with graphical representations, sequence editing, sequence alignment, restriction enzyme mapping, protein structure visualization, mutation and structure superimposition programs along with phylogenetics tree construction supporting dendrograms, neighbor joining and unweighted pair group method with arithmetic mean (UPGMA) programs. Genomic studies like linkage programs are also implemented. Special emphasis has been paid to integrate all the resources in one software so that the researcher does not have to install numerous pieces of software to analyze his data.

Key words: Bioinformatics, linkage, visualizer, alignment, superimposition, phylogenetics.

INTRODUCTION

BioMatriX is developed in Biojava language and major help and support is taken from Biojava Cook book (http://www.biojava.org/wiki/BioJava: CookBook3.0). BioMatriX is an interactive, multi-functional and user friendly bioinformatics tool kit. It represents similarities with most of the famous scientific research work benches like ExPAsy Proteomics Server (http://expasy.org/) and CLC Bio (http://www.clcbio.com/). BioMatriX is a desktop application just like CLC Bio and composite of various modules and functionalities implemented with graphical outputs in order to facilitate research analysis from various aspects. Although it has adopted many features of ExPAsy and CLC Bio (http://www.clcbio.com/ index.php?id=30), an effort has been made to represent sequences with proper scaling and color scheme. The module of genetic linkage is a distinguishing feature of BioMatriX which is not found integrated in any workbench yet. There are many standalone protein visualizers Like RASMOL (http://www.umass.edu/microbio/rasmol/) (Rodger and James, 1995) that are freely available, but BioMatriX itself has a visualizer and other structural manipulation functions implemented in it. It is freely available for download for academic use for the scientific community (http://www.bmx-biomatrix.blogspot.com). An overview of the software main interface is shown in Figure 1.

SOFTWARE PROGRAMS

BioMatriX supports multiple file format reader, format converter and other file writing manipulation functions. This extensive tool kit comprises the following various modules implementing several programs.

DNA sequence analysis

This module is especially designed for DNA/RNA sequence analysis, which include basic functions like calculating nucleotide composition, DNA complement, DNA reverse complement, DNA transcription, RNA com-

*Corresponding author. E-mail: shagufta.kanwal@iiu.edu.pk.

plement, RNA reverse complement, protein translation, open reading frames (ORF) finding and alternate protein translation. This module also supports graphical representations like nucleotide concentration plot and molecular weight plot and nucleotide composition plot as shown in Figure 2.

Another useful feature of this module is sequence editing, which provides various sequence manipulation functions like sequence search by selecting sequence location and range, sequence insertion at the any selected location, sequence deletion by selecting location and number of nucleotides to be deleted like frameshift, point mutation, missense/non-synonymous and nonsense mutations. Sequence alignment provides pair-wise alignment by two methods: the local alignment by implementing Smith-Waterman Algorithm (Smith and Waterman, 1981), and the global alignment by Needleman-Wunsch Algorithm (Zhihua and Lin, 2004). Multiple sequence alignment integrates ClustalW program (Thompson et al., 1994) which is a well known frequently used alignment tool with reliable results. Dot plot has been implemented for quick comparative visualization of two sequences with possible optimal matches in diagonal direction.

Protein sequence analysis

This module is designed for protein sequence analysis such as amino acid composition, amino acid to nucleotide conversion, molecular weight, charge density (PI) and protein nature (acidic/basic) along with graphical representations. The protein polarity plot, molecular weight plot, protein flexibility plot, accessibility plot, antigenic plot, exposed plot, turn plot, hydrophilicity plot and hydrophobicity plotting on several scales like Engleman-Steitz, Hopp-Woods, Kyte-Doolittle, Janin, Chothia & Eisenberg-Weiss, as evaluated by Kallol et al. (2003) are also implemented in this module (http://www.clcbio.com/sciencearticles/BE-hydrophobicity.pdf). This module also supports sequence editing functions like sequence search, insertion, deletion and mutation. Likewise, sequence alignment is also implemented as earlier mentioned in the DNA sequence analysis.

Phylogenetic analysis

This module constructs phylogenetic trees. Outputs of multiple sequence alignment are taken as input to this program to display dendrograms. One of the examples of alignment tree is mentioned in Figure 3. Apart from displaying dendrograms, two important methods of constructing phylogenetics trees are implemented: the neighbor joining (Saitou and Nei, 1987) and UPGMA (Backeljau et al., 1996). These programs take distance matrix as input to display neighbor joining and the unweighted pair group method with arithmetic mean (UPGMA) trees.

Structure analysis

This module provides protein structure analysis- extracts PDB information like structure ID, chains, length and residues information. Another program, Mutate a residue, is also implemented in order to mutate an amino acid and to show the change in protein structure. As an example, one of proteins, 1TNF has been displayed in Figure 4 with its original structure in one window and its structure after mutation is displayed in another window for comparative structural analysis.

One of the useful features of this module is structure visualizer which is named as BIOMOL by integrating Jmol (http://www.jmol.org). This tool acquires all basic structure manipulations and visualization functions as that of JMOL. Another important feature of this module is structure superimposition. This program displays two different protein structures in two separate windows and superimposed structure in the third window along with alignment in the output tab.

Genetic linkage analysis

This module is subdivided into calculating linkage at single point locus and multipoint locus calculation. Single point locus helps in creating the pedigree file which is used as an input to the linkage program in order to find recombinants and non recombinants of family data of any genetic disorder and LOD score calculation on it. This module also displays graphical representations of LOD score plot and recombinant/non-recombinant ratio plot. An example of pedigree data and LOD score calculation are shown in Figure 5. Second is multipoint locus calculation which includes preparation of PRE, PED and DAT files required by MLINK (http://hg.wustl.edu/info/linkage/mlink.html) which is a free ware linkage program (Goldgar and Oniki, 1992) and integrated in it. This module is integrated with a database of genetic markers in order to facilitate a research scientist or geneticist to keep the records of the genetic markers and corresponding reports. This database helps in adding new records, updating the old ones, deleting records and searching for genetic markers via markers identifier number (ID). Apart from analysis modules, various other miscellaneous programs have been implemented like restriction enzyme cutter and file format converter.

CONCLUSION

BioMatriX offers services with multiple functions on the researcher's desk and freely available for the scientific community. It is designed in an easy to use and environment friendly manner with production of precise results in minimum time. The use of this application will prove to be helpful in major analysis like sequence and structure of biomolecules (DNA/RNA/Protein), comparative, evolutionary and genetic studies. Therefore, it will play an important role to build, mine and explore the scientific knowledge and can be a vital entity of every researcher at his desktop.

REFERENCES

Backeljau T, Bruyn LD, Jordaens LDWK, Dongen SV, Winnepenninckx TB (1996). Multiple UPGMA and Neighbor-joining Trees and the Performance of Some Computer Packages. Mol. Biol. Evol. 13(2): 309-313.

Goldgar DE, Oniki RS (1992). Comparison of a multipoint identity-by-descent method with parametric multipoint linkage analysis for mapping quantitative traits. Am. J. Hum. Gene. 50(3): 598-606.

Kallol MB, Daniel RD, John GD (2003). Evaluation of methods for measuring amino acid hydrophobicities and interactions. J. Chromatogr. A. 1000: 637-655

Rodger S, James EM (1995). RasMol: Biomolecular graphics for all, Trends in Biochem. Sci. (TIBS*).* 20: p. 374.

Saitou N, Nei M (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. 4(4): 406-425.

Smith TF, Waterman MS (1981). Identification of Common Molecular Subsequences. J. Mol. Biol. 47: 195-197.

Zhihua DU, Lin F (2004). Improvement of the Needleman-Wunsch Algorithm. SpringerLink, 3066: 792-797.

Thompson JD, Higgins DG, Gibson TJ (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 22(22): 4673-4680.