A NEW ENTERPRISE SCALE SOFTWARE SYSTEM FOR THE ANALYSIS OF THE BIOLOGICAL DATA: AN ENTERPRISE LIFEWARE

Ahmed K. Atwa[1], Aboelenine S. Aboelenine[1], Mai S. Mabrouk[2], Yasser M. Kadah[1]
[1]Institute of Biomedical Engineering, Cairo University, Giza, Egypt
[2]Department of Biomedical Engineering, Misr University, 6October, Egypt
e-mail: Eng_Ahmed_Kam@ieee.org

Abstract--**In this paper, we propose a new software system design for the analysis and processing of the biological data, namely the sequence data and the microarray gene expression profiles. The efficiency and reliability requirements of the biological signal processing environment along with the interoperability and portability of the components constitute the system have imposed a limitation on the selection of the software tools and the underlying architecture of the system. We have also employed a number of state-of-the-art statistical signal processing algorithms to rather discover the underlying biological processes in the microarray gene expression experiments. On the other hand, due to the overwhelming nature of the biological signals it was also required to build a grid of loosely coupled software services to carry out the most computationally intensive bioinformatics algorithms. Moreover, we have also implemented a web interface to facilitate the use of our system.**

Keywords **- bioinformatics, computational biology, service-oriented architecture (SOA), grid-computing, DNA sequences, microarray processing, biotechnology**

I. INTRODUCTION

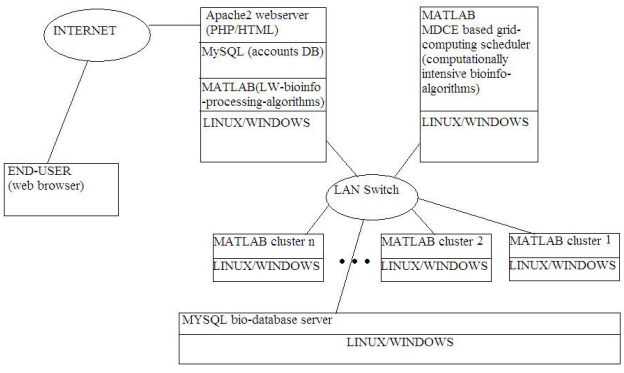
The ever increasing biological data driven by the significant technological advancements in the field of biotechnology is fueling the need for bioinformatics tools to analyze, process, store the biological data, and more importantly, to extract interesting biological patterns. These tools play important roles in enhancing the efficiency of the drug discovery process and even in the emerging field of synthetic biology. Basically we will consider two types of biological data, the sequence data, and the microarray gene expression profile. However, the goal of our project is to produce cross-platform n-tier enterprise scale service oriented architecture (SOA) that provides a comprehensive biological processing framework and an intuitive web user interface. Throughout the development of the project, we have adopted the rational-unified-process (RUP) as it provides an iterative risk-driven software development methodology. Due to the requirements of our software, we have considered a number of candidate software tools to implement both of the web interface and the algorithmic infrastructure. Nevertheless, we have developed a Linux/Windows-Apache2-Mysql-PHP-Matlab based system since it provides an optimal mix of software tools given the requirements and the time constraints. The algorithmic base is entirely built in MATLAB including the fine grained message-passing-interface (MPI2) grid while the web interface was built mainly in PHP since it is scalable and simple. However, to address some interoperability issues relating PHP and MATLAB we have introduced a new concept of a file (I/O) based semaphores so that it provides a cross-platform implementation and on the other hand, it will not create a bottleneck in the process since the computational intensive bioinformatics algorithms will dominate the execution time in all cases. Furthermore, in our project we have used the standard HTTP protocol running on top of TCP/IP protocols to link the network's components together. The implemented bioinformatics modules can be categorized according to the nature of the bio-data into two groups, sequence analysis modules and microarray gene expression profile analysis modules. The first group encompasses the following functionality: pair-wise sequence alignment, grid based multiple sequence alignment, phylogenetic analysis, and sequence statistical analysis while the second group encompasses the following functionality: gene expression normalization and visualization, gene filtration, unsupervised gene clustering, and PCA/ICA based gene grouping. A typical scenario of the running system is as following: 1- a remote user equipped with a web browser logs into the system. 2- The user starts up a bio-module and submits the input data which can be either an uploaded local file or a primary key of some public biological database over the internet. 3- The user gets the result.

In this paper, we demonstrate the capabilities of our software system emphasizing the employed algorithms to accomplish both of the biological sequences processing and the microarray gene expression processing.

II. METHODOLOGY

The implementation of our software system involves a MATLAB based bioinformatics algorithmic framework, a set of software interfaces to enable the system's components to be interoperable with each other, an intuitive web based user interface.

1) *The system's architecture*: Throughout the project's inception phase, we have benchmarked a large set of software tools and programming languages in order to estimate the optimal combination of software tools that satisfies our functional-usability-reliability-performance-supportability-interface-security (FURPS+) requirements. An apache2-Mysql-PHP-MATLAB based system built on top of Linux and/or windows platform has delivered the optimal results in our case.



The scalability of PHP along with the efficiency of MATLAB have made the set of MATLAB-PHP to a large extent an optimized web based bioinformatics computational environment that also provides an easier and more productive programming capabilities when compared to the other cross-platform alternatives for e.g. CGI/Fast-CGI/JSP/Servlets based scripts and the varieties of the bioinformatics software tools that are developed on the other strongly and weakly typed languages (bio-Perl, bio-Java, bio-Python…etc) and even R. However, due to some interoperability issues between MATLAB and PHP, we have developed a file (I/O) based semaphores in order to synchronize their execution concurrently. A MYSQL database system was employed to enable our system to store the biological data and to log the output results for later retrieving and/or processing. That is, it functions as a Web cache as well as a data logger. Therefore, our system can be thought of as an HTTP based web-service that employs a bio-database, a matlab distributed computing engine based fine grained (MPI2) grid-computing capability, and a large scale framework for the biological data processing.

2) *DNA sequence analysis modules*: The input to these modules is a set of DNA sequences (whether from a public internet database or an uploaded local files) and it is required to compare (align) them, find their corresponding evolutionary relationship (phylogenetic tree), or perform statistical analysis (amino-acid/dimmer/codon counting, logo estimation) on them. However, we have employed the following techniques in order to accomplish these tasks optimally as shown in the table below.

Sequence analysis module	Employed technique/algorithm
Global and local pair-wise sequence alignment	Dynamic programming based Needleman-Wunsch and Smith-Waterman algorithms along a statistical randomization testing algorithm to find which ORFs are significantly aligned using blosum, dayhoff, gonnet, and the pam scoring matrices. It also displays the basic dot-plot of the two sequences' ORFs.
Multiple sequence alignment	Progressive dynamic programming based multiple alignment.
phylogenetic tree estimation	Neighbor joining method based on the pair-wise distances using the gonnet scoring scheme.
Sequence statistical analysis and logo finding	Basic counting and string processing algorithms and an entropy based logo finder.

Nevertheless, we have implemented a transparent data acquisition system that is able to identify the sequences formats autonomously and is capable of fetching the sequences from the internet's public databases. That is, it supports the following formats and databases: EMBL, FASTA, Genbank, and Genpept; it also supports the files in the affymetrix format.

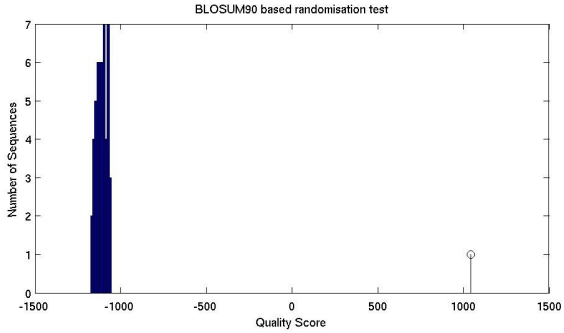
3) *Microarray gene expression analysis modules*: In order to extract interesting biological information from the simultaneous expression levels of the genes (the green and the red channels) we have employed a number of statistical analysis and processing techniques. That is, we have implemented a set of modules in order to normalize, visualize, filter, cluster, perform principal-component-analysis (PCA) based grouping, and perform independent-component-analysis (ICA) on the uploaded microarray file in either of the following formats: spot, genepix, affymetrix, and mat (matlab data file). The following table indicates the employed techniques to accomplish the microarray data analysis.

Microarray analysis module	Employed technique/algorithm
Microarray normalization and visualization	Gene expression mean column based normalization, median filtration, foreground and background spatial visualization in two different color maps (normal, hot), box plot, intensity ratio scatter plot, loglog plot.
Microarray filtration	Small-profile (row) variance filtration, small profile range filtration, low-absolute values filtration, low entropy based filtration
Microarray unsupervised clustering	K-means and hierarchical clustering and dendrogram profiler algorithms
Microarray genes grouping	Principal component analysis (PCA) and PC based grouping (clustering).
Microarray ICA analysis	FastICA based Independent components estimation.

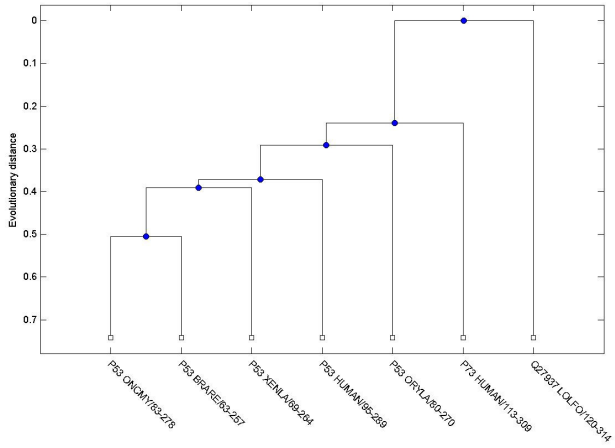
III. Results

A. DNA sequence analysis modules

We have tested our software using real DNA sequences datasets. That used datasets were in the kilobases scale. Our software system is designed, nevertheless, to handle the genome scale sequences as well since it provides a scalable grid that can expand to any number of clusters beyond our 4 clusters based grid. However, we have compared the human Hexa gene open-reading-frames (ORFs) and the mouse Hexa gene ORFs though our pair-wise-sequence-alignment module. The third ORF of the human Hexa has produced a significant alignment with the first ORF in the mouse Hexa gene. The following figure depicts the Blosum90 based randomization test result of the alignment.

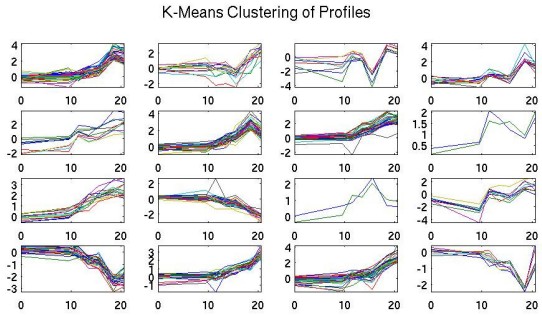


Furthermore, we have tested a set of seven cellular tumor antigen (p53) DNA sequences through our grid based multiple sequence alignment and phylogenetic (evolutionary) analysis as shown in the figure below.

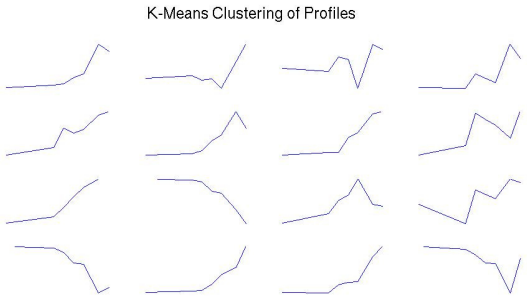


B. Microarray gene expression analysis modules

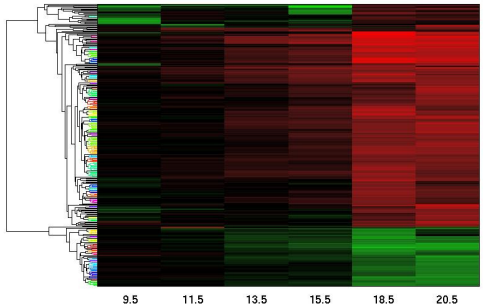
A 6400 relative gene expression levels (log2 ratio of channel1's mean and channel2's mean) of Saccharomyces Cerevisiae during the metabolic shift from fermentation to respiration were analyzed. These relative expression levels were measured at seven time points during the diauxic shift. 310 genes were significant in the experiment after the filtration step. The following figure illustrates the output of the K-means based gene clustering module of these genes.



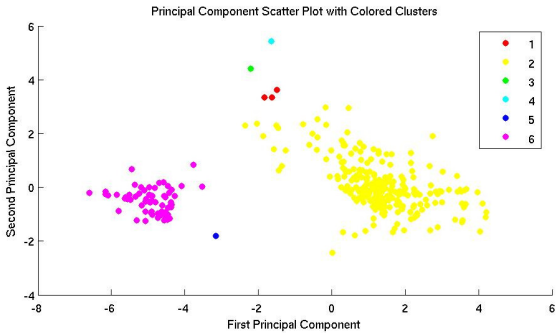
The centroids of these profiles are also computed as shown in the figure below.



Moreover, in order to reveal the hierarchical relationships between the given genes we have plotted their heat map and dendrogram as following:



Furthermore, using our gene-grouping module we have clustered the genes based on their first two principal components. The following figure shows colored clusters of genes based on this technique.



The genes names can also be retrieved from the grouping module's web page.

IV. Discussion and Conclusion

In this paper, we have presented a new software system for the analysis of two types of the biological data namely the DNA sequences and the microarray gene expression profiles. We have selected a set of software tools to fit our productivity oriented vision of the system. The efficient bio-processing framework provided by our system along with its parallel processing capability has proven to be rather encouraging to adopt our proposed software architecture. That is, the ever increasing efficiency of today's weakly typed languages particularly MATLAB is widening the applications of these languages especially in the productivity oriented bioinformatics field. We, also, have introduced a set of bio-processing algorithms that are available through our software system. Among these algorithms, grid-based multiple sequence alignment and phylogenetic (evolutionary) analysis, microarray visualization, microarray genes clustering using different techniques.

In conclusion, we would like to emphasize that, our approach to the bioinformatics software development has proven to be rather encouraging not only in the bioinformatics field but also to any other computationally intensive endeavor.

References

[1] Pierre Bladi, and Soren Brunak. "Bioinformatics: the machine learning approach," *The MIT press. Cambridge, MA, 1st edition*, February 1998.

[2] D. M. Mount, "Bioinformatics: sequence and genome analysis," *Cold spring harbor laboratory press. Cold spring harbor NY, 2nd edition*, July 2004.

[3] V.M. Brown, A. Ossadtchi, A. H. Khan, S. Yee, G. Lacan, W. P. Melega, S. R. Cherry, R. M. Leahy, and D. J. Smith, "Multiplex three dimensional brain gene expression mapping in a mouse model of a Parkinson's disease," Genome research 12(6): 868-884, 2002.

[4] J. L. De Risi, V. R. Iyer,, and P.O. Brown. "Exploring the metabolic and genetic control of gene expression on a genomic scale," Science, 278(5338):680-6, October 1997.

[5] A. Wozniak. "Using video oriented instructions to speed up sequence comparison," Oxford Journal on Bioinformatics, Vol.13 No.2 pp.145-150, November1996.

[6] Te-Won Lee. "Independent component analysis: theory and applications," *Kluwer academic press. Boston, MA, 1st edition*, 1998.

[7] A Hyvarinen and E. Oja, "A fast fixed point algorithm for the independent component analysis," Neural computation,, 1997,9, pp. 1483-1492.

[8] N. Saitou, and M. Nei, "*The neighbor-joining method: a new method for reconstructing phylogenetic trees",* 3 rd ed., Molecular biology and evolution, 1987, 4(4):406-25.

[9] Duan C. Hanselman, and Bruce L. Littlefield. "Mastering MATLAB 7," *Prentice hall press. MA, 1st edition*, October 2004.

[10] Philippe Kruchten. "The rational unified process: an interoduction," Addison-Wesley professional. *MA, 1st edition*, December 2003.