International Journal of Biometrics and Bioinformatics (IJBB)

ISSN : 1985-2347

VOLUME 4, ISSUE 6

PUBLICATION FREQUENCY: 6 ISSUES PER YEAR

International Journal of Biometrics and Bioinformatics (IJBB)

Volume 4, Issue 6, 2011

Edited By
Computer Science Journals
www.cscjournals.org

Editor in Chief Professor João Manuel R. S. Tavares

International Journal of Biometrics and Bioinformatics (IJBB)

Book: 2011 Volume 4, Issue 6
Publishing Date: 08-02-2011
Proceedings
ISSN (Online): 1985-2347

IJBB Journal is a part of CSC Publishers
http://www.cscjournals.org

Published in Malaysia

Typesetting: Camera-ready by author, data conversation by CSC Publishing Services – CSC Journals, Malaysia

CSC Publishers

Editorial Preface

This is the fifth issue of volume four of International Journal of Biometric and Bioinformatics (IJBB). The Journal is published bi-monthly, with papers being peer reviewed to high international standards. The International Journal of Biometric and Bioinformatics are not limited to a specific aspect of Biology but it is devoted to the publication of high quality papers on all division of Bio in general. IJBB intends to disseminate knowledge in the various disciplines of the Biometric field from theoretical, practical and analytical research to physical implications and theoretical or quantitative discussion intended for academic and industrial progress. In order to position IJBB as one of the good journal on Bio-sciences, a group of highly valuable scholars are serving on the editorial board. The International Editorial Board ensures that significant developments in Biometrics from around the world are reflected in the Journal. Some important topics covers by journal are Bio-grid, biomedical image processing (fusion), Computational structural biology, Molecular sequence analysis, Genetic algorithms etc.

The coverage of the journal includes all new theoretical and experimental findings in the fields of Biometrics which enhance the knowledge of scientist, industrials, researchers and all those persons who are coupled with Bioscience field. IJBB objective is to publish articles that are not only technically proficient but also contains information and ideas of fresh interest for International readership. IJBB aims to handle submissions courteously and promptly. IJBB objectives are to promote and extend the use of all methods in the principal disciplines of Bioscience.

IJBB editors understand that how much it is important for authors and researchers to have their work published with a minimum delay after submission of their papers. They also strongly believe that the direct communication between the editors and authors are important for the welfare, quality and wellbeing of the Journal and its readers. Therefore, all activities from paper submission to paper publication are controlled through electronic systems that include electronic submission, editorial panel and review system that ensures rapid decision with least delays in the publication processes.

To build its international reputation, we are disseminating the publication information through Google Books, Google Scholar, Directory of Open Access Journals (DOAJ), Open J Gate, ScientificCommons, Docstoc and many more. Our International Editors are working on establishing ISI listing and a good impact factor for IJBB. We would like to remind you that the success of our journal depends directly on the number of quality articles submitted for review. Accordingly, we would like to request your participation by submitting quality manuscripts for review and encouraging your colleagues to submit quality manuscripts for review. One of the great benefits we can

provide to our prospective authors is the mentoring nature of our review process. IJBB provides authors with high quality, helpful reviews that are shaped to assist authors in improving their manuscripts.

Editorial Board Members
International Journal of Biometrics and Bioinformatics (IJBB)

Editorial Board

Editor-in-Chief (EiC)

Associate Editors (AEiCs)

Editorial Board Members (EBMs)

Table of Content

Volume 4, Issue 6, December 2011

Pages

Real Time Web-based Data Monitoring and Manipulation System to Improve Translational Research Quality

Matthew N. Anyanwu manyanwu@uthsc.edu
Clinical and Translational Science Institute
University of Tennessee Health Science Center
Memphis, TN

Venkateswara Ra Nagisetty nnagise@uthsc.edu
Clinical and Translational Science Institute
University of Tennessee Health Science Center
Memphis, TN

Emin Kuscu ekuscu@uthsc.edu
Clinical and Translational Science Institute
University of Tennessee Health Science Center
Memphis, TN

Teeradache Viangteeravat tviangte@uthsc.edu
Clinical and Translational Science Institute
University of Tennessee Health Science Center
Memphis, TN

Abstract

The use of the internet technology and web browser capabilities of the internet has provided researchers/scientists with many advantages, which includes but not limited to ease of access, platform independence of computer systems, relatively low cost of web access etc. Hence online collaboration like social networks and information/data exchange among individuals and organizations can now be done seamlessly. In practice, many investigators rely heavily on different data modalities for studying and analyzing their research/study and also for producing quality reports. The lack of coherency and inconsistencies in data sets can dramatically reduce the quality of research data. Thus to prevent loss of data quality and value and provide the needed functionality of data, we have proposed a novel approach as an ad-hoc component for data monitoring and manipulation called RTWebDMM (Real-Time Web-based Data Monitoring and Manipulation) system to improve the quality of translational research data. The RTWebDMM is proposed as an auditor, monitor, and explorer for improving the way in which investigators access and interact with the data sets in real-time using a web browser. The performance of the proposed approach was evaluated with different data sets from various studies. It is demonstrated that the approach yields very promising results for data quality improvement while leveraging on a web-enabled environment.

Keywords: Bioinformatics, Health Management, Clinical Trial, Basic Research, Data Manipulation, Data Monitoring, Data Cleaning, Data Comparison

1. INTRODUCTION

A data with good quality is needed in-order to produce high quality results from scientific researches and discoveries. This has generated a considerable amount of interest in software/algorithms that can facilitate data quality control. The value of data highly depends on its quality. To enhance the quality of data to be analyzed many data management systems tend to facilitate data quality control before using it in data mart, data mining, or other analytical processes. Data quality control includes all the processes involved in producing and validating good quality data. The processes include but not limited to data-preprocessing/cleaning, data processing, data aggregation and data quality assurance [11]. Data preprocessing involves noise and dimension reduction in the data. In Data processing stage, data is analyzed, aggregated, and incorrect data items eliminated. Also data is examined in this stage for reasonable output [11]. Data aggregation is a measure of the statistical test and analysis of the processed data. Also different statistical tests used in analyzing the data out-put are validated.

Data quality assurance involves the use of quality assurance techniques/methods to validate the data output. Data quality control and validation is used to ensure that good and authoritative data is produced for its purpose [12]. Fan et al., proposed the Semandaq (Semantic Data Quality) [9]. Semandaq is a data quality system that uses conditional functional dependencies for improving the quality of relational data. Galhardas et al., proposed a declarative language along with 1-5 transformation operations to enhance improvements of data monitoring and cleaning process [3]. Harris et at., proposed the Research Electronic Data Capture (REDCap) [4]. The REDCap project uses PHP + JavaScript programming language and MySQL database engine driven methodology and workflow process. The REDCap proposed Data Cleaner and Data Comparison tools to assist in data monitor and cleaning process. However the cleaning actions have to be explicit by the investigators or users. Viangteeravat et al., introduced the Scientific Laboratory Information Management–Patient-care Research Information Management (Slim-Prim) system [7, 8].

The Slim-Prim [7,8] proposed Data Monitor tool to assist the user/researcher with real-time visualization and tracking of the historical data set through Asynchronous JavaScript and XML (AJAX) capability interface. However, it gives users the ability to refine and manipulate their data set to support monitoring. Also the cleaning of poor quality data is still needed under the principal best known as “Your Data, Your Decision”. Raman et al., introduced an interactive ad-hoc technique by providing a spreadsheet-like interface to facilitate the specific transformation operation that can automatically trigger a bad quality of data in the background [6]. Mury et al., proposed the Informatics for Integrating Biology and the Bedside (i2b2) that queries data by dragging and placing the data items into the query environment. This approach is used in retrieving data item from the repository which contains clinical data records [13]. In practice, however, it is estimated that data cleaning is the most labor intensive and a complex process compared to other processes in data quality control [10]. In order to minimize the data cleaning efforts, data quality control process should be part of the data processing stage as it involves data monitoring and manipulation. This stage produces good errors and detects inconsistencies in data items.

In this article, we have proposed the real-time web-based data monitoring and manipulation system (RTWebDMM) as an ad-hoc component that can easily be integrated and interfaced with an existing data management system while having the advantages of being an online web-based system. RTWebDMM provides a graphical user interface (GUI) platform in form of a spreadsheet with build-in macros analytical tools that perform both data monitoring and manipulation in-order to produce good quality data for research purpose. It has also shown to be an indispensable tool in data cleaning efforts, thereby relieving the researcher the burden of data cleaning.

2. RTWebDMM SYSTEM ARCHITECTURE AND FUNCTIONALITY

The Real-Time Web-based Data Monitoring and Manipulation system (RTWebDMM) architecture is depicted in Figure 1. As shown Figure 1, the RTWebDMM is composed of four main components. The first component, Ad-hoc API, uses PHP + Asynchronous JavaScript and XML(AJAX) and JavaScript programming language to build an ad-hoc API (Application Programming Interface) to communicate with the existing Clinical Data Management System (CDMS) and uploads data set of interest into RTWebDMM. The second component comprises the Data monitoring and Data Manipulation sub-components. The Data monitoring sub-component consists of built-in Macro tools, data analysis tools, Data graph (which produces the graphical user interface) and Data tracking tool while the Data manipulation sub-component consists of Data Export which may be in Comma-Separated Value (CSV) or Tab Separated Value (TSV) format, Data Aggregation and Comment Exchange modules. Data Query module initiates the built-in Macro module to work on analytical process using Data Analysis. The Data Analysis is composed of many built-in complex analytical functions such as Mean, Median, Standard deviation, Age calculation, Probability Density Function, BMI (Body Mass Index), Standard error, and Outlier detection.

The RTWebDMM is based on a Simple Spreadsheet [6, 14]. A Simple Spreadsheet provides a basic excel-like framework that supports charts, formulas, and simple custom macros. It also includes sorting capabilities, which has the features of expanding all or specified columns, complex macro built-in analytic formulas, and other features necessary for basic science and clinical research. We have also extended it to include the organization of what we call “Data Query”. Data Query is the real-time SQL (Structured Query Language) that allows a user to easily access and monitors the quality of the data set. The user can interactively manipulate, track the state/status of data, if it has been modified since it was last collected by using the Data Tracking module. The Data Aggregation gives users the ability to manipulate the raw data set to a suitable format before statistician(s) can provide further analysis on the data. The Data Conversion becomes the third component used in creating custom user report, which is static mode (i.e., data snapshot). The fourth component, Custom Report, is an online report module. The report module is used to know the state of the system at any given time.

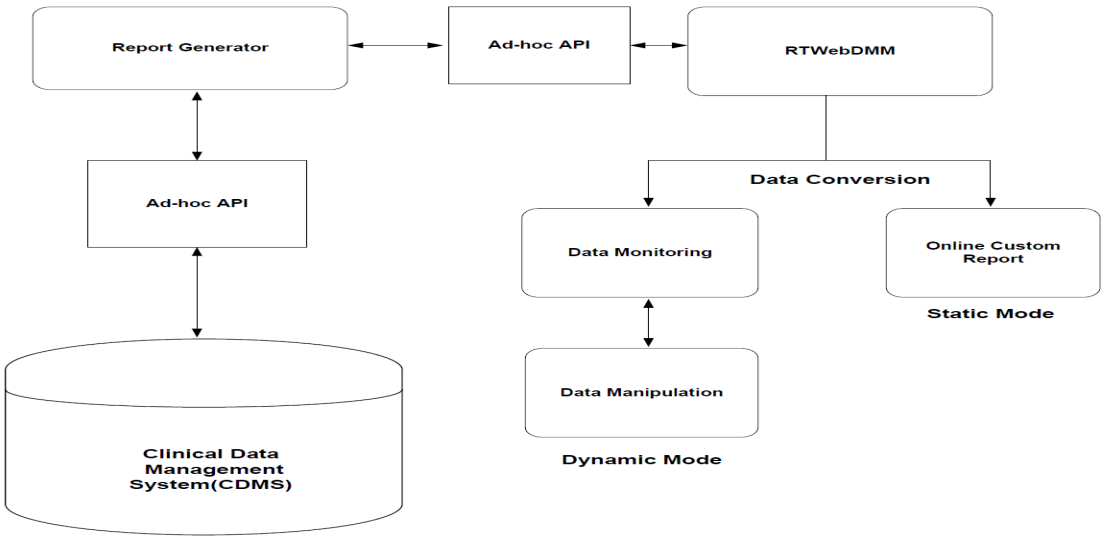

FIGURE 1: RTWebDMM system architecture and functionality

3. RTWebDMM APPLICATION PROGRAMMING INTERFACE (API)

For the purpose of our study, there is an API between our system and CDMS [7, 8] which enables the uploading of data into the RTWebDMM project using the API in Data Query menu as depicted in Figure 2. The RTWebDMM uses PHP application running at the server site to communicate and upload raw data set from Slim-Prim system. The JavaScript (JS) programming language is used to display the query results at the client-side as shown in Figure 3.

In Figure 2, RTWebDMM uses Structure Query Language (SQL) to communicate with the CDMS system. The result of the raw data set is then translated into a required format in which its structure is compliant for rendering to the client using JS Data Grid Writer and Render Class. The JS Data Grid Writer and Render Class is the PHP class written in OOP (Object-Oriented Programming) fashion. It is used to create the compliant JS format for RTWebDMM to render and display its result in dynamic data grid style. Once RTWebDMM successfully renders the displays of its results, also the user is allowed to manipulate and monitor the data set using Data Graph, Data Tracking, and Data Aggregation modules thus detecting outlier or poor quality of data. This process reduces the extensive labor in data manipulation and assists in data cleaning.

The RTWebDMM uses Data Graph and Tracking components to monitor the value of the data set. The user uses either simple built-in formulas (e.g., Mean, Median, or Standard deviation) or complex built-in formula (e.g., Body Mass Index(BMI)) in detecting the quality of data. To enhance collaboration, RTWebDMM provides comment functionality for data comment exchange as shown in Figure 3.

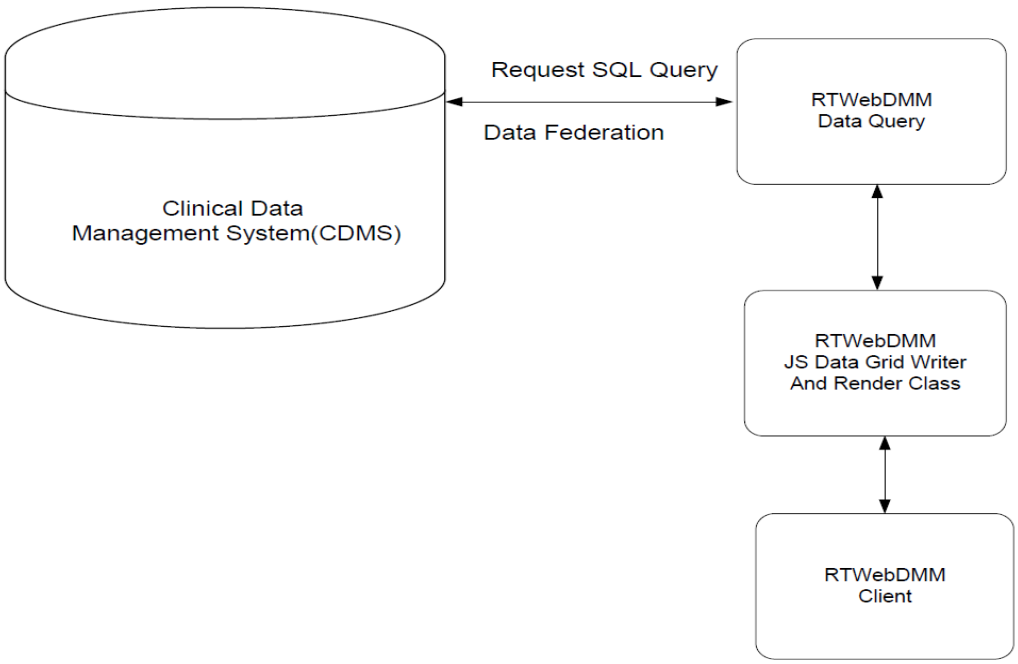

FIGURE 2: RTWebDMM system API (Application Programming Interface)

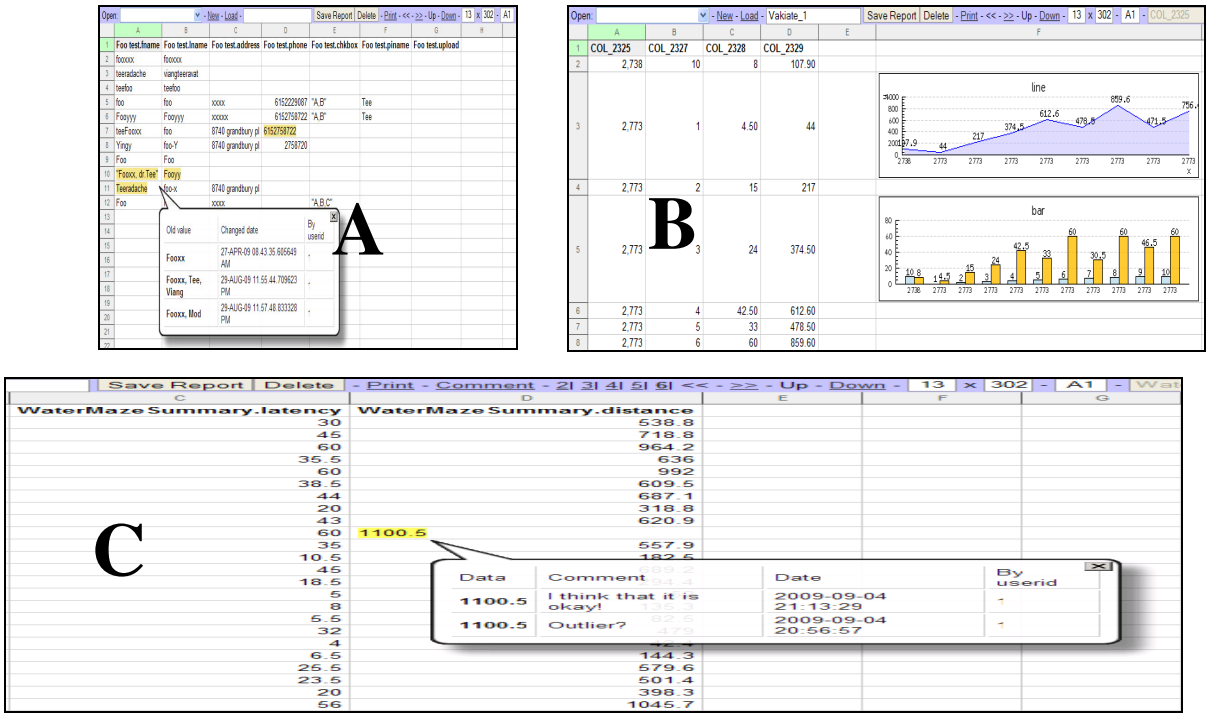

FIGURE 3: RTWebDMM User Interface (web-based) for Historical Data Tracking(A) Data Graph / Analysis(B) and Collaboration supported in comment exchange functionality(C)

4. DISCUSSION AND FUTURE CHALLENGES

The proposed Real-Time Web-based Data Monitoring and Manipulation (RTWebDMM) system has shown significant improvement in reducing extensive labor for data cleaning process. Studies available in literature have shown that it is difficult and challenging to detect poor quality of data conditions in both basic science and clinical research studies [12, 15]. The RTWebDMM attempts to assist in providing a new alternative method in which we are able to monitor uncertainties in relational data sets using built-in independency functions of the data sets. Compared with other data management tools mentioned (See Section 1 and 2) our proposed tool (RTWebDMM) is a real-time web-based data monitoring and manipulation tool and it ensures that good quality data is generated for research purpose. It also has graph Data Graph and Tracking components which monitors the value and quality of data in real-time. RTWebDMM also has API for ease of integration with legacy systems or other data management systems. In practice, basic science and clinical research data deal with relational data dependencies. For instance, the specific zip code within the same county must result in the same name of a city. The ad-hoc functional dependencies that the user can define will have to be established in our future work. The improvements in user-friendly interface (UI) and data cleaning are also part of our future implementation of our system. We also intend to make this tool play a key role in decision support management by creating a repository of commonly used data sets in clinical research. An association between the data items will be created using Apriori Principle which states that “if an item-set is frequent, then all of its subsets must also be frequent” [16]. There will be an alert to signify the presence or absence of a data set item used in the query analysis that has an associate in the repository.

5. CONSLUSION

RTWebDMM tool has been tested with data from clinical research and basic science studies at UTHSC to evaluate its use in improving the quality of data item sets and in eliminating inconsistencies detected in the data sets. It is a web-based tool that gives users of clinical trial research studies real-time access to monitor, manipulate and refine data set items to obtain good quality data for their studies. Data quality affects the result of clinical research studies as the data items are patient medical records which should contain data of highest possible quality that can be obtained. In-order to obtain credible result from clinical research study, the data must be credible [15]. The tool also relieves the user the burden of data cleaning, thus allowing user to focus on the objective of the research study. In future users of clinical research project will leverage on the built-in intelligence of the tool that will be added to determine the association between clinical data set items. The built-in intelligence tools will include the use of standardized techniques and technologies such as; International Classification of Diseases, Ninth Revision (ICD-9) [17], International Classification of Diseases, Tenth Revision (ICD-10) [18], Current Procedural Terminology, 4th Edition (CPT-4) [19], Systematized Nomenclature of Medicine-Clinical Terms (SNOMED CT) [20], Logical Observation Identifiers Names and Codes (LOINC), Recommended Standard Clinical Drug Nomenclature (RxNorm) [21], and Unified Medical Language System (UMLS) [20].

6. REFERENCES

1. C. C. Shilakes and J. Tylman. Enterprise information portals. Technical report, Merrill Lynch, Inc., New York, NY, Nov. 1998.

2. C. Gao, W. Fan, F. Geerts, X. Jia, and S. Ma. Improving data quality: Consistency and accuracy. In VLDB, 2007.

3. H. Galhardas, D. Florescu, D. Shasha, E. Simon, and C.-A. Saita. Declarative data cleaning: Language, model and algorithms. In VLDB, 2001.

4. P. A. Harris, R. Taylor, R. Thielke, J. Payne, N. Gonzales, J. G. Conde. Research electronic data capture (REDCap)-A metadata-driven methodology and workflow process for providing translational research informatics support. Journal of Biomedical Informatics 42 (2009) 377-381.

5. PHP Hypertext Preprocessor. http://www.php.net/. 2009. Ref Type: Electronic Citation.

6. V. Raman and J. M. Hellerstein. Potter's wheel: An interactive data cleaning system. In VLDB, 2001.

7. T. Viangteeravat, I. M. Brooks, W.J. Ketcherside, R. Houmayouni, N. Furlotte, S. Vuthipadadon, & C.S. McDonald. Clinical & Translational Science Biomedical Informatics Unit (BMIU): Slim-Prim system bridges the gap between laboratory discovery and practice, 2009.

8. T. Viangteeravat, I. M. Brooks, E.Smith, N. Furlotte, S. Vuthipadadon, R. Reynolds, & C.S. McDonald. "Slim-Prim: A biomedical informatics database to promote translational research". Perspectives in Health Information Management, 2009.

9. W. Fan, F. Geerts, X. Jia, and A. Kementsietsidis. Conditional functional dependencies for data cleaning. TODS, 33(1), 2008.

10. Press release, Garter, Inc.. Quoting Bill Hostmann, Research Director, presented at Gartner Business Intelligence Summit in London, UK., February 3, 2005.

11. Y, Akiyama, and K. S. K. Propheter, “Methods of Data Quality Control: For Uniform Crime Reporting Programs”. Criminal Justice Information Services Division Federal Bureau of Investigation. April 2005.

12. A. S. Loebl, "An organizational and historical perspective of a decade of data validation R&D at the Oak Ridge reservation". Data quality control theory and pragmatics Pages: 1 - 6, 1991.

13. S. N. Murphy, G. Weber, M. Mendis, V. Gainer, H. Chueh, C.,Churchill, S., et al. Serving the enterprise and beyondwith informatics for integrating biology and the bedside (i2b2).Journal of the American Medical Information Association, 17,124–130. PMID: 20190053. 2010.

14. D. J Power, "A Brief History of Spreadsheets", DSSResources.COM, World Wide Web,http://dssresources.com/history/sshistory.html, version 3.6, 08/30/2004. Photo. Retrieved: September 24, 2010.

15. J. Rothenberg. “A Discussion of Data Quality for Verification, Validation, and Certification (VV&C) of Data to be used in Modeling”. Rand Project Memorandum PM709-DMSO. 1977. P. Tan, M. Steinbach, and V. Kumar. “Introduction to Data Mining”. Addison Wesley, New York, 2006.

16. Centers for Disease Control and Prevention ."International Classification of Diseases,Ninth Revision (ICD-9)”. http://www.cdc.gov/nchs/icd/icd9.htm. Retrieved: January, 2011.

17. Centers for Disease Control and Prevention. "International Classification of Diseases,Ninth Revision (ICD-9)”. http://www.cdc.gov/nchs/icd/icd10.htm. Retrieved: January, 2011.

18. American Medical Association. "CPT” http://www.ama-assn.org/ama/pub/physician-resources/solutions-managing-your-practice/coding-billing-insurance/cpt/about-cpt.shtml . Retrieved: January, 2011.

19. United States National Library of Medicine. "SNOMED Clinical Terms” http://www.nlm.nih.gov/research/umls/Snomed/snomed_main.html. Retrieved: January, 2011.

20. C. McDonald , S.M. Huff, J. Suico , & K.Mercer (eds). Logical Observation Identifiers Names and Codes (LOINCR) users’ guide. Indianapolis: Regenstrief Institute; 2004.

Biological Significance of Gene Expression Data using Similarity based Biclustering Algorithm

J.Bagyamani bagya.gac@gmail.com
Government Arts College,
Dharmapuri - 636705,
TamilNadu, India

K. Thangavel drktvelu@yahoo.com
Department of Computer Science,
Periyar University, Salem - 636 011,
TamilNadu, India

R. Rathipriya rathipriyar@yahoo.co.in
Department of Computer Science,
Periyar University Salem - 636 011,
TamilNadu, India

Abstract

Unlocking the complexity of a living organism's biological processes, functions and genetic network is vital in learning how to improve the health of humankind. Genetic analysis, especially biclustering, is a significant step in this process. Though many biclustering methods exist, only few provide a query based approach for biologists to search the biclusters which contain a certain gene of interest. This proposed query based biclustering algorithm SIMBIC+ first identifies a functionally rich query gene. After identifying the query gene, sets of genes including query gene that show coherent expression patterns across subsets of experimental conditions is identified. It performs simultaneous clustering on both row and column dimension to extract biclusters using Top down approach. Since it uses novel 'ratio' based similarity measure, biclusters with more coherence and with more biological meaning are identified. SIMBIC+ uses score based approach with an aim of maximizing the similarity of the bicluster. Contribution entropy based condition selection and multiple row / column deletion methods are used to reduce the complexity of the algorithm to identify biclusters with maximum similarity value. Experiments are conducted on Yeast Saccharomyces dataset and the biclusters obtained are compared with biclusters of popular MSB (Maximum Similarity Bicluster) algorithm. The biological significance of the biclusters obtained by the proposed algorithm and MSB are compared and the comparison proves that SIMBIC+ identifies biclusters with more significant GO (Gene Ontology).

Keywords: Data Mining, Bioinformatics, Biclustering, Gene Expression Data, Gene Selection, Top-Down Approach, Gene Ontology**.**

1. INTRODUCTION

Gene expression is conversion of information encoded in a gene. Gene expression data is a valuable resource for researchers who are focusing on clustering of genes to draw meaningful

inferences. Expressions of genes under different conditions serve as valuable clues to understand the cell differentiation, pathological and genetic behavior. For most functionally related genes, tight correlation occurs under specific experimental conditions. Clustering deals with finding patterns in a collection of unlabeled data. Traditional clustering algorithms consider all of the dimensions of an input dataset in an attempt to learn as much as possible about each object described. According to Kerr et. al [12], clustering the microarray matrix can be achieved in two ways: (i) genes can form a group which show similar expression across conditions, (ii) samples can form a group which show similarity across all genes. This gives rise to *global clustering* or *traditional clustering* where a gene or sample is *grouped across all dimensions*. Biclustering [15, 21], a relatively new unsupervised learning technique, cluster the objects under subset of attributes. It allows the assignment of individual objects to multiple clusters. Co-expressed genes, i.e., genes with similar expression patterns, can be clustered together and manifest similar cellular functions. Hence biclustering aims to find sub-matrices with coexpressed expression values.

1.1 Query driven Biclustering

In this Query driven Biclustering technique, usually a query gene is given as input, and a single bicluster which consists of a set of genes and a subset of conditions / samples that are similar to the query gene is extracted. The resultant bicluster that include the query gene answer the following questions which are not answered by most existing biclustering methods in which biologists are interested in [7].

(i) "Which genes involved in a specific protein complex is co expressed?"
(ii) "Given a set of known disease genes, how to select new candidate genes that may be linked to the same disease?"

Given a specific gene or set of genes (seed genes) known or expected to be related to some common biological pathway or function:

(i) "Which genes are (functionally) related to the seed genes and which features (conditions) are relevant for this biological function?"

1.2 Biological Significance

An **Open Reading Frame** (ORF) is a DNA sequence that contains a start codon and a stop codon in the same reading frame. ORF is supposed to be a gene which encodes a protein, but in some cases encoded protein for ORFs are not known. The yeast *Saccharomyces cerevisiae* [13] is an excellent organism for this type of experiment because its genome has been sequenced and all of the ORFs have been determined. Each study determines the expression level of every ORF at a series of time points. The resulting dataset must be analyzed to determine the roles of specific genes in the process of interest. Genes coding for elements of a protein complex are likely to have similar expression patterns. Hence, grouping ORFs with similar expression levels can reveal the function of previously uncharacterized genes.

1.3 Coherent Bicluster

Genes involved in common processes are often co-expressed. In this paper, constant bicluster with reference to the query gene and coherent bicluster with reference to the query gene are extracted. The biological significance of both the biclusters with reference to the same query gene is identified. Comparison of the biological significance shows that coherent bicluster has more biological significance than the constant bicluster. Hence the focus in identifying coherent (i.e., patterns that rise and fall concordantly) bicluster is that co-expression may reveal much about the genes' regulatory systems. Coherent bicluster [1] has more biological significance than constant bicluster.

1	2	5	0
2	3	6	1
4	5	8	3
5	6	9	4
Additive Coherent Bicluster			

1	2	0.5	1.5
2	4	1	3
4	8	2	6
3	6	1.5	4.5
Multiplicative coherent Bicluster			

TABLE 1: Additive Coherent Bicluster and Multiplicative coherent Bicluster

This paper is organized as follows: Section 2 details the preliminary of gene expression data along with literature survey. Section 3 explains the proposed work and the evaluation measures. Section 4 provides the experimental results of Yeast Saccharomyces Cervisiae expression data. Biological validation of the genes within the bicluster is provided in terms of gene ontology in Section 5. Section 6 concludes the article.

2. Background

2.1 Microarray Gene Expression Data

Genes are how living organisms inherit features from their ancestors. The information within a particular gene is not always exactly the same between one organism and another, so different copies of a gene do not always give exactly the same instructions. Gene expression levels can be determined for samples taken (i) at multiple time instants of a biological process (different phases of cell division) or (ii) under various conditions (e.g., tumor samples with different histopathological diagnosis). A gene expression database can be regarded as consisting of three parts – the gene expression data matrix, gene annotation and sample / condition annotation.

2.2 Problem statement

A gene expression matrix A = [a_{ij}] of size m x n where each element represents the expression level of gene 'i' under condition 'j' is considered. Let I be the set of genes and J the set of conditions of A. Biclustering identification is to find a submatrix $A_{I'J'}$ = A (I', J') with sets of rows I'$\subset$ I and sets of columns J' $\subset$ J . . In general, the problem can be defined as one of finding large sets of rows and columns such that the rows show unusual similarities along the dimensions characterized by columns and vice-versa. The bicluster cardinality or volume of bicluster is simply the product of the number of genes and number of conditions in the bicluster.

2.3 Nature of biclustering Algorithms

Biclustering, which has been applied intensively in molecular biology research recently, provides a framework for finding hidden substructures in large high dimensional matrices Tanay et al. [19, 20] defined a bicluster as a subset of genes that jointly respond upon a subset of conditions. Biclustering algorithms may have two different objectives: to identify one bicluster or to identify a given number of biclusters. This proposed method identifies *one bicluster at a time*.

Many biclustering methods [3] such as iterative row column [6,8] divide and conquer [9], exhaustive bicluster enumeration, distribution parameter identification exist in literature. Greedy iterative search methods are based on the idea of creating biclusters by adding or removing rows/columns from them, that optimizes the given criteria. They may make wrong decisions and loose good biclusters, but they have the potential to be very fast.

Cheng and Church [5] used a greedy procedure starting from the entire data matrix and successively removing columns or rows contributing most to the mean squared residue score. They used both single node deletion and multiple node deletion methods in order to arrive one bicluster at a time and mask the previously discovered biclusters. Iterative Signature Algorithm (ISA) by Ihmels et al [11] has been found to be very effective in identifying (Transcription Module) TMs in yeast expression data. However, the major problem with the algorithm is that it starts with a totally random input gene seed and hence can result in non-meaningful TMs. Thus to gain confidence in the quality of TMs they run their algorithm for a large number of seeds and report a

TM only if it is obtained. Dhollander et al. [7] introduced a model-based query-driven module discovery tool QDB, but it is aimed at performing informed biclustering instead of pattern matching, and it does not take into account the complex correlation patterns such as inverse patterns. Owen et al. [16] proposed a score-based search algorithm called Gene Recommender (GR) to find genes that are co expressed with a given set of genes using data from large microarray datasets. GR first selects a subset of experiments in which the query genes are most strongly co-regulated. Hence multiple query genes are required. Hu et al. [10] developed model-based gene expression query algorithm BEST (Bayesian Expression Search Tool) built under the Bayesian model selection framework. It is capable of detecting co-expression profiles under a subset of samples/experimental conditions. In MSB [14] the maximum similarity bicluster for query gene or reference gene i* is computed, by trying the algorithm for all the conditions j* and then identifying bicluster with maximum similarity. The advantage of MSB is that it is unnecessary to mask previously discovered biclusters. SIMBIC [2] algorithm is an improvement of MSB in terms of computational efficiency but the biclusters obtained by both the methods are same. Instead of single row / column deletion, multiple rows / columns are deleted. Also for a specific reference gene i*, the algorithm need not be executed for all the reference condition j* but j* can be restricted to n/2 conditions that has high contribution entropy. This proposed SIMBIC+ algorithm is an improved version of SIMBIC in the sense that it uses novel 'ratio' based similarity measure, applied on conditions with high contribution entropy. Also multiple rows or multiple columns are deleted in each iteration until the gene expression matrix reduces to a single element. Then bicluster with maximum similarity is identified and evaluated using ACV (Average Correlation Variation) measure. The biological significance and p - value of each obtained bicluster are evaluated. The Gene Ontology (GO) of the biclusters obtained by the proposed SIMBIC+ and MSB are compared and the comparison shows that SIMBIC+ outperforms SIMBIC and MSB.

3. PROPOSED WORK

3.1 Condition selection

Preprocessing often involves some operation on feature-space in order to reduce the dimensionality of the data. This is referred to as feature selection [17]. The features are sorted based on the contribution entropy value. SVD-based entropy [18] of the dataset is defined as follows. Let s_j denote the singular values of the matrix A. s_j^2 are then the eigen values of the n x n matrix AA^T. The values are normalized by using (1).

$$V_j = s_j^2 / \sum_k s_k^2 \quad (1)$$

and the resulting dataset entropy is

$$E = \frac{1}{\log(N)} \sum_{j=1}^{j=N} V_j \log(V_j) \quad (2)$$

where N is the total number of attributes. This entropy varies between 0 and 1. The minimal value E = 0 corresponds to an ultra ordered dataset and E = 1 corresponds to unordered dataset. The contribution of the i^{th} feature to the entropy CE_i is defined by a leave-one-out comparison according to

$$CE_i = E(A_{[n \times m]}) - E(A_{[n \times (m-1)]}) \quad (3)$$

where, in the last matrix, the i^{th} feature is removed. Thus the features are sorted by their relative contribution to the entropy. Simple ranking (SR) method sorts the features. Select 'n/2' features / conditions according to the highest ranking order of their CE_i values.

3.2 Ratio based Similarity between genes

Gene selection is critical in molecular class prediction. In a cellular process, only a relatively small set of genes are active. So select genes i* which has specific functional importance in gene

ontology viz. Cellular component, Biological process, Molecular function. Let i* be a reference gene / query gene.
Let j* be the reference condition. j* may be chosen in such a way that it has high contribution entropy. The contribution entropy of all the conditions are computed and j* is chosen from the selected 'n/2' conditions of the expression data that has high contribution entropy. Because there is a dependency between co-expression and functional relation, co-expressed genes provide excellent candidates for further study. However, the dependency is complex, and it cannot be used to identify the best choice of similarity measure. In [2, 14], the similarity measure is based on the absolute value of the difference. This measure would help us to identify constant and additive biclusters. In order to identify a coherent pattern (shifting and scaling pattern), similarity measure is defined in terms of ratio.

For an element a_{ij} of expression matrix A (I, J) and a reference gene i* ∈ I,

$d_{ij} = abs (a_{ij} / a_{i*j})$ and $d_{avg} = \sum_{i \in I} \sum_{j \in J} \frac{d_{ij}}{|I| * |J|}$

where | . | refers to number of elements. The similarity between two genes s_{ij} is defined as

$$s_{ij} = \begin{cases} 0 & if \ d_{ij} \geq d_{avg} \\ 1 - \frac{d_{ij}}{d_{avg}} & otherwise \end{cases} \quad (4)$$

If $d_{ij} > d_{avg}$, then the two elements a_{ij} and a_{i*j} are not similar and the similarity s_{ij} is set to 0.

3.3 Ratio based Similarity score for a bicluster

Let S (I, J) be an m x n similarity matrix of A (I, J). The similarity score S (I, J) of the bicluster A_{IJ} is defined as below.

For row i ∈ I, the similarity score of row 'i' is S (i, J) = $\sum_{j \in J} s_{ij}$ (5)

For row j ∈ J', the similarity score of column 'j' is S (I, j) = $\sum_{i \in I'} s_{ij}$ (6)

The similarity score of bicluster S (I, J) = min {min S (i, J), min S (I, j)} (7)

If this minimum is min(S(i, J)) find the index of all the rows corresponding to this minimum and remove **all** those rows from A(I, J) to get A(I', J) else find the index of the columns corresponding to column minimum and remove **all** those columns from A(I, J) to get A(I, J'). Then A (I, J) is updated as A (I, J') or A (I', J). Multiple row / column deletion is performed until the the row size (mr) or column size (mc) is less than or equal to 1. Identify the bicluster which has high similarity score as maximum similarity bicluster. Popular measures used for evaluating quality of a bicluster are MSR (Mean Squared Residue)[5] and ACV (Average Correlation Variation)[4] measure. MSR measures well all types of constant biclusters [1] and ACV is perfect measure for coherent biclusters.

SIMBIC+ Algorithm

Constant bicluster:

Input

1. Gene expression matrix A(I, J)
2. Reference gene i* which has GO functional importance.
3. Reference condition j* from selected (n/2) features.

Output a maximum similarity bicluster.

Procedure

1. Compute similarity matrix S (I, J) using (4) for the reference gene i*.

2. Parameters (mr, mc) = size (A (I, J)).
3. While (mr ≤ 1 or mc ≤ 1)
4. Compute row_sim, S(i, J) $= \sum_{j \in J} s_{ij}$
5. Compute col_sim, S(I, j) = $\sum_{i \in I} s_{ij}$
6. $\forall i$, find min(S(i, J) and $\forall j$, find min S(I, j)
7. Find min { min(S(i, J') & min S(I', j)}
8. If this minimum is min(S(i, J')) find the index of the rows corresponding to this minimum and remove all those rows from A(I , J) to get A(I', J)
9. else find the index of the columns corresponding to column minimum and remove all those columns from A(I , J) to get A(I , J').
10. Update A(I, J) = A(I', J) or A(I, J) = A(I, J') and S(I, J) = S(I', J) or S(I, J) = S(I, J')
11. Find the similarity of bicluster using (7) for the updated S (I, J).
12. Update mr, mc.
13. End while
14. Extract the bicluster with maximum similarity A (I', J').
15. Compute ACV and MSR of A (I', J').

3.4 Comparison of SIMBIC+ with MSB

MSB	SIMBIC+
Every row is considered as a reference gene i*.	Only genes with functional importance are considered as reference gene i*
Every column is considered as a reference column j*.	The (n/2) conditions that have more contribution entropy are considered as j*.
Number of iterations is m+n-2.	Number of iterations is very less.
Single node deletion method is used.	Multiple node deletion method is used.
Distance measure is the absolute difference between the reference gene and other genes.	Distance measure is the ratio between the reference gene and other genes.
Similarity measure depends on the parameters α and β.	No such parameters used for bicluster identification.
More complex.	Complexity and number of iterations are reduced.
Biclusters have biological significance.	Biclusters have still more biological significance.

TABLE 2: Comparison of MSB and SIMBIC+

This SIMBIC+ algorithm is implemented in Matlab, 2GHz processor with 3 GB RAM.

4. Experimental analysis

4.1 Dataset

In order to test the efficiency of the proposed algorithm the Yeast Saccharomyces Cerevisiae data with 2884 genes and 17 conditions was considered wherein the missing values are replaced by -1. [http://arep.med.harvard.edu/biclustering/]

4.2 Bicluster Evaluation Measures

Two types of biclusters namely constant and additive coherent are identified using this algorithm. It is observed from Table 3 that additive biclusters have more biological significance than the constant biclusters. The performance of the algorithm is validated using MSR and the ACV. For each bicluster, MSR and ACV are computed using the formulae

$$MSR = \sum_i \sum_j r_{ij}^2 \tag{8}$$

where $r_{ij} = a_{ij} - \mu_{ik} - \mu_{jk} + \mu_k$, μ_{ik} is the row mean, μ_{jk} is the column mean and μ_k is the mean of the bicluster.

$$ACV = max \left\{ \sum_{i=1}^{i=m} \sum_{j=1}^{m} \frac{|c\_row_{ij}|}{m^2 - m}, \sum_{p=1}^{n} \sum_{q=1}^{n} \frac{|c\_col_{pq}|}{n^2 - n} \right\} \tag{9}$$

where c_row_{ij} is the correlation coefficient between rows i and j and c_col_{pq} is the correlation coefficient between columns p and q. Bicluster with low MSR and high ACV (i.e., ACV approaching 1) is a good bicluster. 'P' value of a bicluster provides the biological significance of a bicluster. It provides the probability of including genes of a given category in a cluster by chance. Thus overrepresented bicluster is a cluster of genes which is very unlikely to be obtained randomly. Suppose that we have a total population of *N* genes, in which *M* have a particular annotation. If we observe *x* genes with that annotation, in a sample of *n* genes, then we can calculate the probability of that observation, using the hyper geometric distribution. Thus the probability of getting *x* or more genes with an annotation, out of *n*, given that *M* in the population of *N* have that annotation, is:

$$p\_value = 1 - \sum_{j=0}^{x-1} \frac{\binom{M}{j}\binom{N-M}{n-j}}{\binom{N}{n}} \tag{10}$$

The gene ontology namely Biological Process (BP), Molecular Function (MF) and Cellular Component (CC) of the bicluster can be identified using **GOTermfinder**.

4.3 Performance of SIMBIC+ Algorithm

Table 3 gives the comparison of the performance of the proposed algorithm for corresponding reference gene i* and reference condition j* for identifying a maximum similarity bicluster of Yeast Saccharomyces Cerevisiae dataset. It is observed that the first four biclusters of Table 3 identified by the proposed SIMBIC+ are highly correlated compared to bicluster obtained from MSB for the same reference gene and reference condition. Even though the last two biclusters of Table 3 identified by MSB are more correlated (with high ACV) the volume of the bicluster is comparatively less i.e., statistically these are good biclusters. Statistical significance alone does not decide the quality of the bicluster. Statistical measures evaluate a bicluster theoretically, but the biological significance proves the real quality of the bicluster obtained. Hence the biological significance of the biclusters obtained by the proposed SIMBIC+ and MSB are tabulated in Table 4 and Table 5 respectively.

i*	j*	Nature of bicluster	SIMBIC+			MSB		
			No. of Iterations	ACV	Size of bicluster	No. of Iterations	ACV	Size of bicluster
210	14	Constant	1903	**0.4864**	20 x 17	2899	0.3165	25 x 17
210	14	Additive	2647	**0.9553**	18 x 16	2899	0.7020	15 x 12
288	14	Constant	1903	**0.3556**	22 x 17	2899	0.2519	22 x 16
288	14	Additive	2583	**0.9684**	19 x 16	2899	0.9224	19 x 14
2462	9	Additive	1759	0.9300	**19 x 17**	2899	**0.9988**	29 x 8
1459	17	Additive	2455	0.9199	**19 x 16**	2899	**1.0000**	6 x 6

TABLE 3: Comparison of performance of SIMBIC+ with MSB

The selected conditions of yeast Saccharomyces data based on the contribution entropy are 6, 7, 8, 9, 12, 13, 14, 15 and 17. Bicluster plots or parallel coordinate plot and heatmaps provide the

visual representation of the bicluster. Figures 1, 3, 5, 7 are the bicluster plots of biclusters obtained by the proposed SIMBIC+ algorithm and Figures 2, 4, 6, 8 are the bicluster plots of biclusters obtained by MSB. Figure 1 is the bicluster plot of additive bicluster with 19 genes, 16 conditions when i* is chosen as 288 (gene ID 'YBR198C' which has the functional importance of SLIK (SAGA like complex) and reference condition j* is chosen as 14. This bicluster has ACV = 0.9684 and MSR= 9.7747 x 10^4. Figure 2 is the bicluster plot of additive bicluster with 19 genes, 14 conditions for the same reference gene and reference condition. This bicluster has ACV = 0.9224 and MSR = 5.3994 x 10^4. Figure 3 shows the bicluster plot of additive bicluster with 19 genes and 16 conditions when i* is chosen as 210 and reference condition j* is chosen as 14. This bicluster has ACV= 0.9553 and MSR= 7.6272 x 10^4. Figure 4 shows the bicluster plot of additive bicluster with 15 genes and 12 conditions for the same reference gene and reference condition. This bicluster has ACV= 0.7020 and MSR= 4.6092 x 10^4.

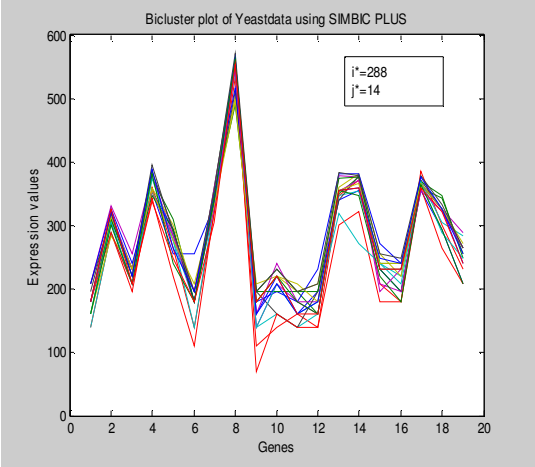

FIGURE 1: Additive Bicluster using SIMBIC+ with i*=288

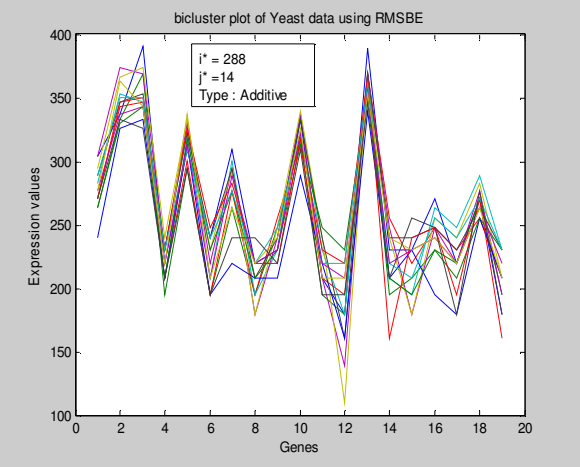

FIGURE 2: Additive Bicluster using MSB with i*=288

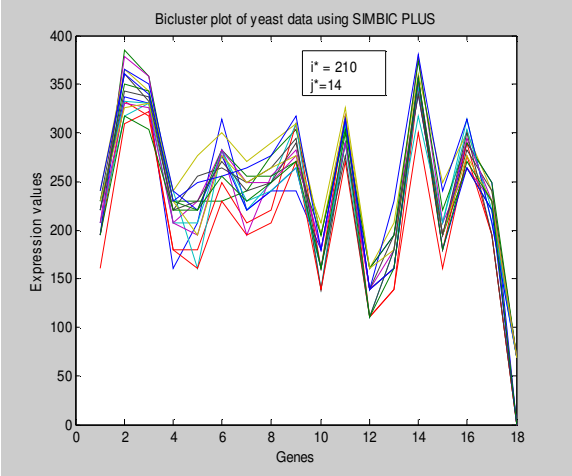

FIGURE 3: Additive Bicluster using SIMBIC+ with i*=210 and j*=14

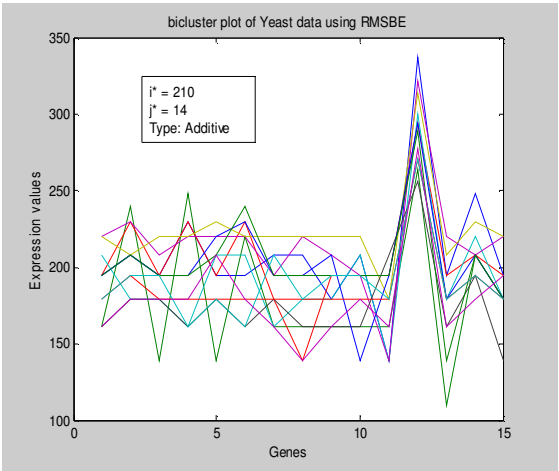

FIGURE 4: Additive Bicluster using MSB with i*=210 and j*=14

Figure5 shows the bicluster plot constant bicluster with 22 genes and 17 conditions when i* is chosen as 288 and reference condition j* is chosen as 14. This bicluster has ACV= 0.3556 and MSR= 1.0717 x 10^5. Figure 6 shows the bicluster plot of constant bicluster with 22 genes and 16

conditions for the same reference gene and reference condition. This bicluster has ACV= 0.2519 and MSR= 8.8503 x 10^4.

Figure7 shows the bicluster plot of constant bicluster with 20 genes and 17 conditions when i* is chosen as 210 and reference condition j* is chosen as 14. This bicluster has ACV = 0.4864 and MSR = 9.9778 x 10^4. Figure 8 shows the bicluster plot of constant with 25 genes, 17 conditions for the same reference gene and reference condition. This bicluster has ACV= 0.3165 and MSR= 1.204 x 10^5.

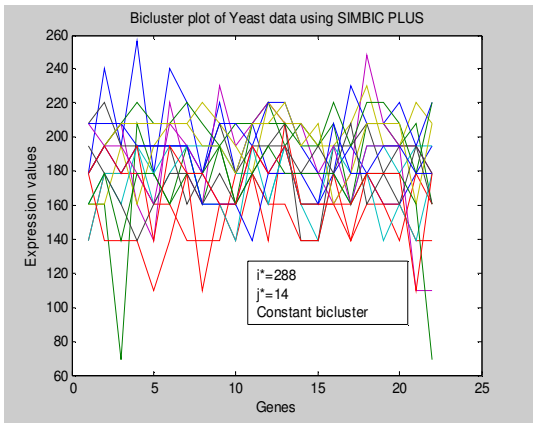

FIGURE 5: Constant Bicluster using SIMBIC+ with i*=288 and j*=14

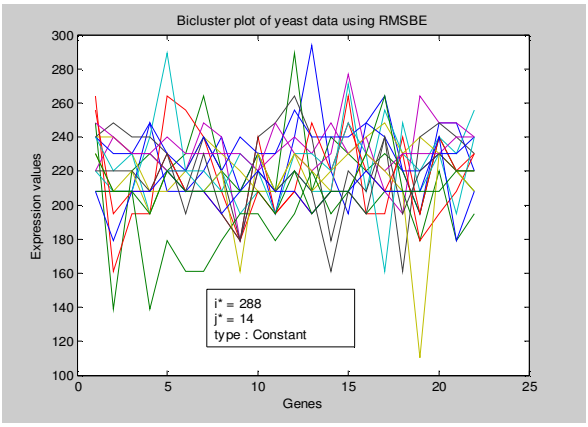

FIGURE 6: Constant Bicluster using MSB with i*=288 and j*=14

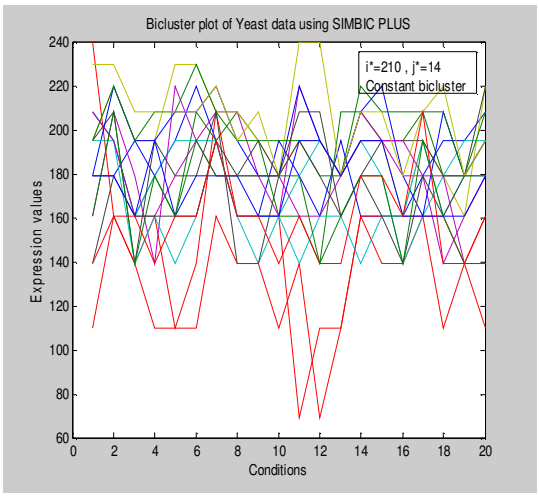

FIGURE 7: Constant Bicluster using SIMBIC+ with i*=210 and j*=14

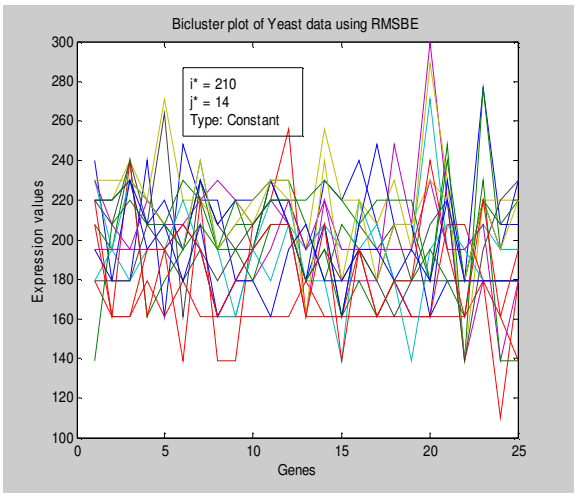

FIGURE 8: Constant Bicluster using MSB with i*=210 and j*=14

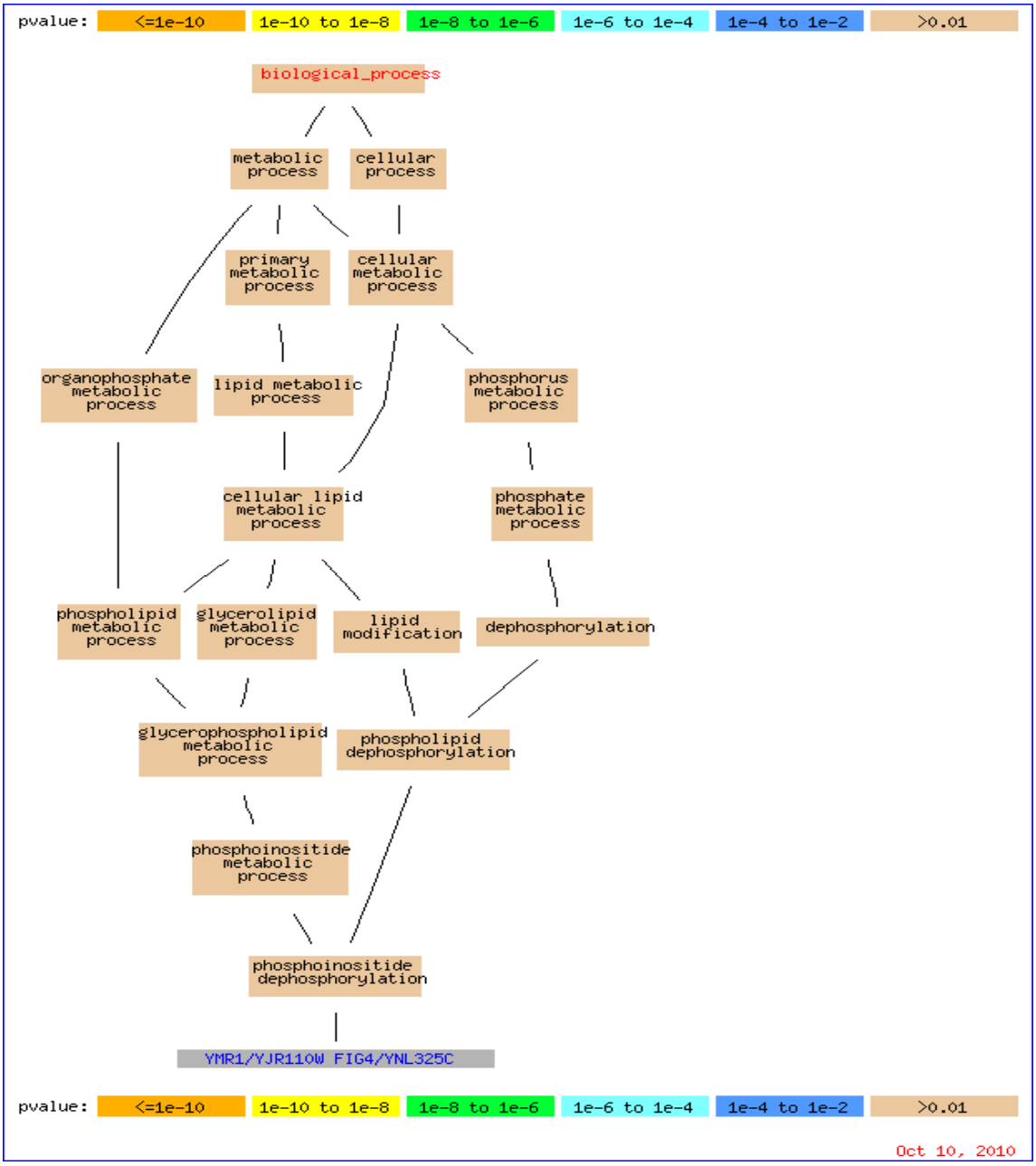

FIGURE 9: Biological significance of constant bicluster with i* = 210 and j*=14 using SIMBIC+

5. BIOLOGICAL VALIDATION

The annotations consist of three ontologies, namely biological process, cellular component and molecular function. The biological significance and the p value are obtained from **GO TermFinder**[1]. From Table 4 and Table 5, it is also observed that bicluster of the proposed SIMBIC+ algorithm are GO enriched. Table 6, provides the comparison of GO of the proposed SIMBIC+ algorithm and GO of MSB algorithm. Also Figures 9, 10 and 11 provide the biological network of the resultant bicluster. Figure 9 provides the GO for constant bicluster of SIMBIC+ with

[1] http://www.yeastgenome.org/cgi-bin/GO/goTermFinder.pl

i* = 210. The genes involved in this bicluster are responsible for biological processes phospholipid dephosphorylation and phosphoinositide dephosphorylation. Figure 10 provides the GO (cellular function) for additive bicluster of SIMBIC+ with i*=210 and j*=14. Figure 11 provide the GO (molecular function) for additive bicluster of SIMBIC+ with i*=288 and j*=14. The genes involved in this bicluster are responsible for ATPase activity, coupled to transmembrane movement of ions, phosphorylative mechanism.

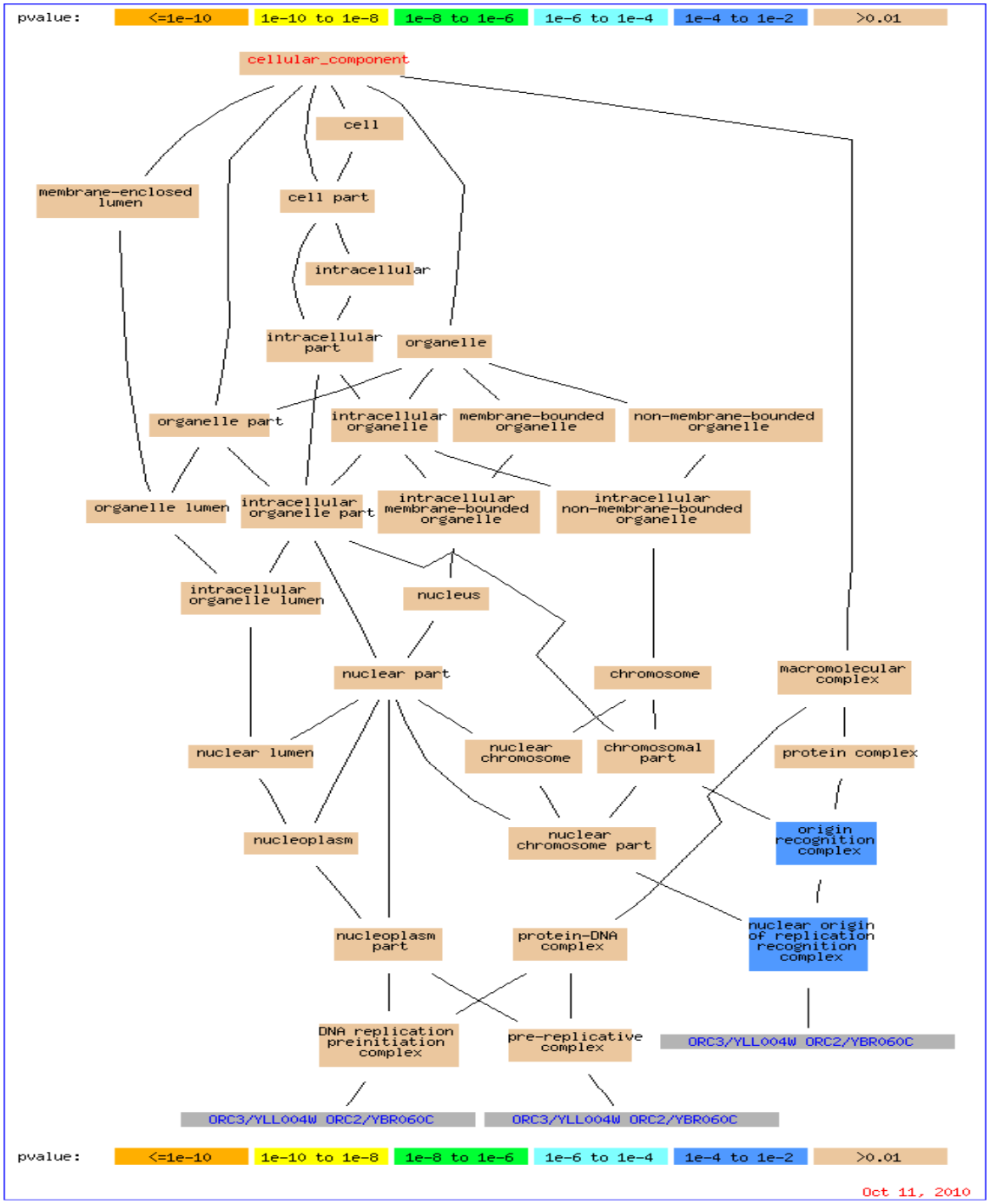

FIGURE 10: Biological significance of additive bicluster i* = 210 and j*=14 using SIMBIC+

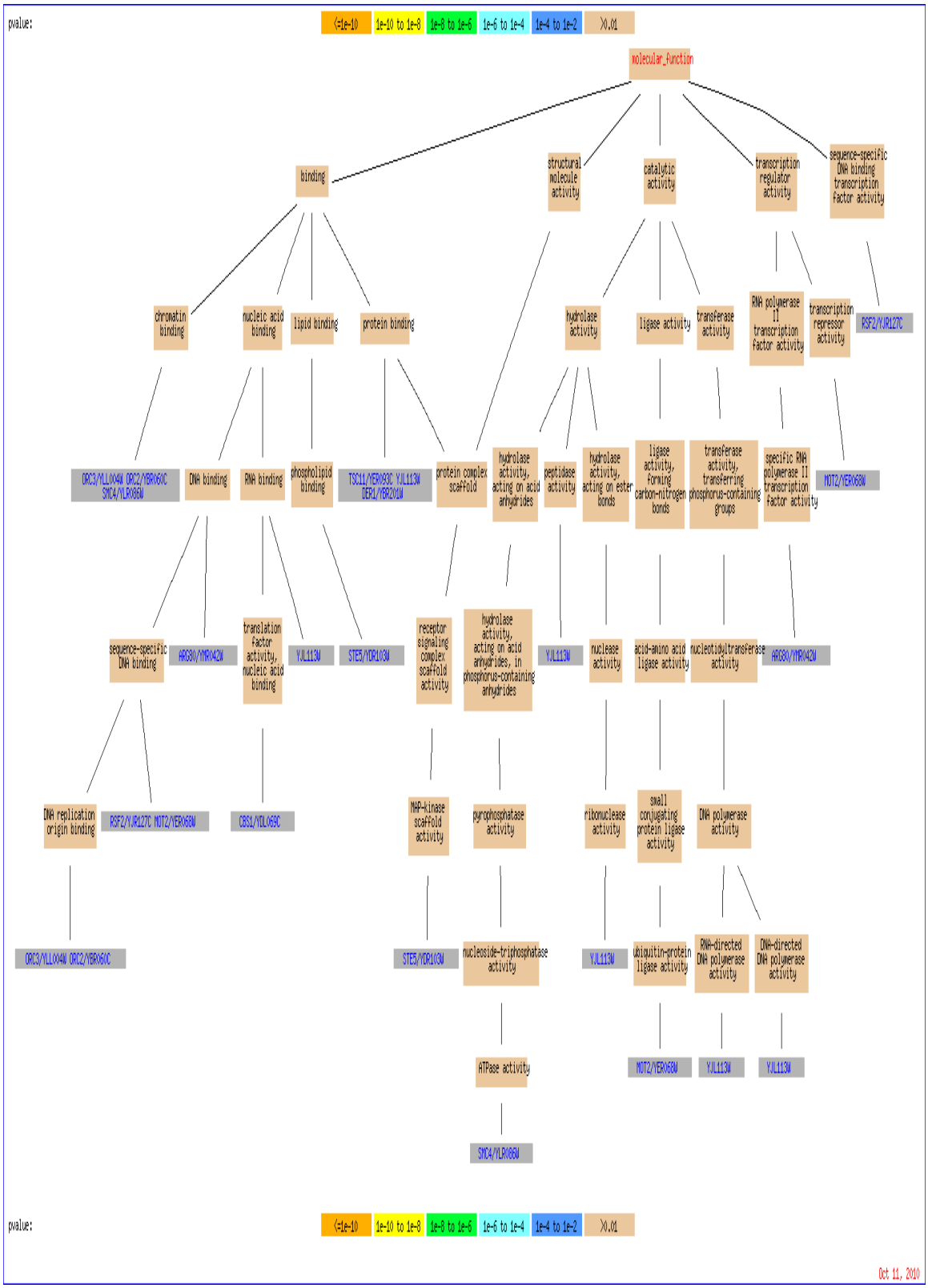

FIGURE 11: Biological significance of additive bicluster i* = 288 and j*=14 using SIMBIC+

Reference gene i*=210 , reference condition j*=14 , alpha=.2 beta = .2 gamma=.9, volume= 15x 12 MSR 46092 ACV=0.7020 , Type : Additive bicluster, GO: Biological Process				
GOID	**GO_term**	**Cluster frequency**	**P-value**	**FDR**
19236	response to pheromone	3 out of 15 genes, 20.0%	0.09218	0.22
Nature of GO: Molecular Function				
4519	endonuclease activity	2 out of 15 genes, 13.3%	0.04723	0.6
Nature of GO: Cellular component unknown				
Reference gene i*=210 , reference condition j*=14 , alpha=.2 beta = .2 gamma=.9, volume= 25x 17 MSR 120400 ACV=0.3165 , Type :Constant bicluster.				
Biological Process - Unknown				
Molecular Function - Unknown				
Cellular component - Unknown				

TABLE 4: Biological significance of Biclusters of Yeast Dataset obtained from MSB

Reference gene i*= 210, Reference condition j* = 14 , volume = 20 *17 = 340 ,MSR =98960 ACV=.4953 Type: constant Bicluster Nature of GO: Biological Process				
GOID	**GO_term**	**Cluster frequency**	**P-value**	**FDR**
46839	phospholipid dephosphorylation	2 out of 20 genes, 10.0%	0.02953	0.18
46856	phosphoinositide dephosphorylation	2 out of 20 genes, 10.0%	0.02953	0.09
9987	cellular process	20 out of 20 genes, 100.0%	0.06939	0.09
Nature of GO: Molecular Function				
3682	chromatin binding	4 out of 20 genes, 20.0%	0.00084	0
Nature of GO: Cellular Component				
4437	inositol or phosphatidylinositol phosphatase activity	2 out of 20 genes, 10.0%	0.00723	0.02
Reference gene i*= 210, Reference condition j* = 14 , volume = 18 *16 = 288 ,MSR =76272 , ACV=.9553 Type: Additive Bicluster Nature of GO:Cellular Component				
Nature of GO: Biological Process				
6814	sodium ion transport	2 out of 18 genes, 11.1%	0.00848	0
15672	monovalent inorganic cation transport	3 out of 18 genes, 16.7%	0.00902	0
Nature of GO: Molecular Function				
15662	ATPase activity, coupled to transmembrane movement of ions, phosphorylative mechanism	2 out of 18 genes, 11.1%	0.01372	0.12
42625	ATPase activity, coupled to transmembrane movement of ions	2 out of 18 genes, 11.1%	0.06971	0.19
44451	nucleoplasm part	6 out of 20 genes, 30.0%	0.00237	0
5654	nucleoplasm	6 out of 20 genes, 30.0%	0.00394	0
43234	protein complex	11 out of 20 genes, 55.0%	0.00874	0
44428	nuclear part	10 out of 20 genes, 50.0%	0.02298	0
46695	SLIK (SAGA-like) complex	2 out of 20 genes, 10.0%	0.04023	0.02
44422	organelle part	14 out of 20 genes, 70.0%	0.04762	0.01
44446	intracellular organelle part	14 out of 20 genes, 70.0%	0.04762	0.01
124	SAGA complex	2 out of 20 genes, 10.0%	0.05593	0.01
70461	SAGA-type complex	2 out of 20 genes, 10.0%	0.06171	0.02
32991	macromolecular complex	12 out of 20 genes, 60.0%	0.08059	0.01

TABLE 5: Biological significance of Biclusters of Yeast Dataset obtained from SIMBIC+

Table:4 provides the biological significance constant and additive biclusters of yeast data for the reference gene i*=210. Table:5 provides the biological significance constant and additive biclusters of yeast data for the reference gene i*=210. There are 2 biological significances for MSB and 19 biological significances for SIMBIC+.Table:6 provides the comparison of GO enrichment of Biclusters of Yeast Dataset obtained by proposed SIMBIC+ and existing MSB algorithms. It is observed that highly correlated biclusters have more biological significance than biclusters with similar values. Also the proposed SIMBIC+ algorithm identifies biclusters with more biological significance (with low 'p' value and less False Discovery Rate).

i*	j*	Type	SIMBIC +			MSB		
210	-	Constant	**3**	**1**	**1**	0	0	0
210	14	Additive	**2**	**2**	**10**	1	1	0
2462	9	Additive	**5**	**3**	1	2	1	**2**
1459	17	Additive	**4**	**2**	**6**	1	1	3
288	14	Constant	**2**	**1**	**3**	2	1	3
288	14	Additive	**3**	**2**	**5**	2	2	4

TABLE :6 Comparison of GO enrichment of Biclusters of Yeast Dataset obtained by SIMBIC+ and MSB

6. CONCLUSION AND FUTURE WORK

This proposed algorithm identifies biclusters of gene expression data with more biological significance. The multiple node deletion method based on the new similarity score applied on the extracted features / conditions, makes the algorithm very efficient and less time consuming. The biological significance of the biclusters and 'p' value are obtained using **GO-Term Finder**. Results prove that the proposed SIMBIC+ algorithm is computationally efficient and biologically significant. Also the results prove that biclusters with scaling pattern are more biologically significant than the biclusters with shifting pattern.

Acknowledgement

The first author acknowledges the UGC, SERO, Hyderabad to carry out this research under FIP. The second author acknowledges the UGC, New Delhi for financial assistance under major research project grant No. F-34-105/2008.

7. REFERENCES

1. W. Ayadi, M. Elloumi, J.K Hao. "A biclustering algorithm based on a Bicluster Enumeration Tree: application to DNA microarray data". Biodata Mining, 2:9, 2009

2. J. Bagyamani, K. Thangavel. "SIMBIC: SIMilarity Based BIClustering of Expression Data". Information Processing and Management Communications in Computer and Information Science, 70, 437-441, 2010

3. A. Ben-Dor, B. Benny Chor, R. Karp, and Z. Yakhini , "Discovering local structure in gene expression data: The order–preserving sub matrix problem". Journal of Computational Biology, 373–84

4. K, Cheng, N. Law, W. Siu and A. Liew. "Identification of coherent patterns in gene expression data using an efficient biclustering algorithm and parallel coordinate visualization" BMC Bioinformatics, 9:210, 2008

5. Y. Cheng, G.M Church, "Biclustering of expression data". Proceedings of 8th International Conference on Intelligent Systems for Molecular Biology, ISMB-00, 93-103, 2000

6. Chun Tang, Li Zhang, Idon Zhang, and Murali Ramanathan, "Interrelated two-way clustering: an unsupervised approach for gene expression data analysis". Proceedings of the 2nd IEEE International Symposium on Bioinformatics and Bioengineering, 41–48, 2001

7. T. Dhollander, Q. Sheng, K. Lemmens, B.D. Moor and K. Marchal et al., "Query-driven module discovery in microarray data". Bioinformatics, 2007

8. G. Getz, E. Levine and E. Domany, "Coupled two-way clustering analysis of gene microarray data". Proceedings of the Natural Academy of Sciences USA, 12079-12084, 2000

9. J.A. Hartigan. "Direct clustering of a data matrix". Journal of the American Statistical Association Statistical Assoc. (JASA), 67, 123-129, 1972

10. M. Hu, and Z.S. Qin. ”Query Large Scale Microarray Compendium Datasets using a Model-Based Bayesian Approach with Variable Selection”, PLoS ONE 4(2) e4495, 2009.

11. J. Ihmels et al. ”Defining transcription modules using large-scale gene expressiondata”. Bioinformatics, 20,2004

12. G. Kerr, H.J. Ruskin, M. Crane and P. Doolan, “Techniques for clustering gene expression data”. Computers in Biology and Medicine, .38 (3), 283-293, 2008

13. J. Laurie Heyer, Semyon Kruglyak, and Shibu Yooseph, “Exploring Expression Data: Identification and Analysis of Coexpressed Genes”. ISMB, Bioinformatics, 22(14), e507-513, 2006

14. X. Liu and L. Wang, “Computing maximum similarity biclusters of gene expression data”, Bioinformatics, 23(1),50-56, 2007

15. S.C. Madeira and A.L Oliveira. “Biclustering algorithms for biological data analysis: a survey”. IEEE Transactions on Computational Biology and Bioinformatics,1(1) 24-45, 2004

16. A.B. Owen, J. Stuart, K. Mach, A.M Villeneuve and S. Kim. “A gene recommender algorithm to identify co expressed genes in C. elegans”. Genome Res 13: 1828–1837, 2003

17. P.M Pardalos, S. Busygin and O.A Prokopyev. “On biclustering with feature selection for microarray data sets”. BIOMAT2005—International Symposium on Mathematical and Computational Biology, World Scientific, 367–78, 2006

18. Roy Varshavsky, Assaf Gottlieb, Michal Linial and David Horn. “Novel Unsupervised Feature Filtering of Biological Data”. Bioinformatics, 22(14), e507-e513, 2006

19. A. Tanay, R. Sharan and R. Shamir. “Biclustering Algorithms: A Survey”. Handbook of Computational Molecular Biology, 2004

20. A. Tanay, R. Sharan and R. Shamir. “Discovering statistically significant biclusters in gene expression data”. Bioinformatics, 18, 136-144, 2002

21. J. Yang, H. Wang, W. Wang and P.S Yu “An improved biclustering method for analyzing gene expression”. International Journal on Artificial Intelligence Tools, 14(5), 771-789, 2005.

A Biological Sequence Compression Based on Cross Chromosomal Similarities Using Variable length LUT

Rajendra Kumar Bharti raj05_kumar@yahoo.co.in
Ph.D. scholar, UTU,
Astt. Professor, CSE Deptt.
B. C. T. Kumaon Engineering College,
Dwarahat, Almora, Uttarakhand, India

Archana Verma vermarchana05@gmail.com
Ph.D. scholar, UTU,
Astt. Professor, CSE Deptt.
B. C. T. Kumaon Engineering College,
Dwarahat, Almora, Uttarakhand, India

Prof. R.K. Singh rksinghkec12@rediffmail.com
Professor, ECE Deptt.
B. C. T. Kumaon Engineering College,
Dwarahat, Almora, Uttarakhand, India

Abstract

While modern hardware can provide vast amounts of inexpensive storage for biological databases, the compression of Biological sequences is still of paramount importance in order to facilitate fast search and retrieval operations through a reduction in disk traffic. This issue becomes even more important in light of the recent increase of very large data sets, such as meta genomes.

The present Biological sequence compression algorithms work by finding similar repeated regions within the Biological sequence and then encode these repeated regions together for compression. The previous research on chromosome sequence similarity reveals that the length of similar repeated regions within one chromosome is about 4.5% of the total sequence length. The compression gain is often not high because of these short lengths of repeated regions. It is well recognized that similarities exist among different regions of chromosome sequences. This implies that similar repeated sequences are found among different regions of chromosome sequences. Here, we apply cross-chromosomal similarity for a Biological sequence compression. The length and location of similar repeated regions among the different Biological sequences are studied. It is found that the average percentage of similar subsequences found between two chromosome sequences is about 10% in which 8% comes from cross-chromosomal prediction and 2% from self-chromosomal prediction. The percentage of similar subsequences is about 18% in which only 1.2% comes from self-chromosomal prediction while the rest is from cross-chromosomal prediction among the different Biological sequences studied. This suggests the significance of cross-chromosomal similarities in addition to self-chromosomal similarities in the Biological sequence compression. An additional 23% of storage space could be reduced on average using self-chromosomal and cross-chromosomal predictions in compressing the different Biological sequences.

Keywords: Biological Sequences, Chromosome, Cross Chromosomal Similarity, Compression Gain, Prediction.

1. INTRODUCTION

The Deoxyribonucleic acid(DNA) constitutes the physical medium in which all properties of living organisms are encoded. Molecular sequence databases (e.g.,EMBL,Genbank, DDJB, Entrez, SwissProt, etc) currently collect hundreds or thousands of sequences of nucleotides and amino acids reaching to thousands of gigabytes and are under continuous expansion. Need for Compression arises because approximately 44,575,745,176 bases in 40,604,319 sequence records are there in the GenBank database[12].

Increasing genome sequence data of organism lead Biological database size two or three times bigger annually[1]. Thus, Efficient compression may also reveal some biological functions and helps in phylogenic tree reconstruction etc[2,3,4]. Compression is desirable to uncover similarities among sequences, and provide a means to understand their properties in addition to reduce storage requirement[13].

There are many text compression algorithms available having quite a high compression ratio[7]. But they have not been proved well for compressing Biological sequences as the algorithm does not incorporate the characteristics of Biological sequences even through sequences can be represented in simple text form. Biological sequences are comprised of just four different bases e.g. for DNA only A,T,G and C[1,2,3,4,7]. Each base can be represented by two bits in binary. So, It has been observed that no file-compression program achieves benchmark of the compression ratio for Biological sequences. Several compression algorithms specialized for Biological sequences have been developed in the last decade and some of these are: Biocompress-2, Gen Compress and CTW+LZ. One knows that all such algorithms take a long time (essentially a quadratic time search or even more) but at the same time achieving high speed and best compression ratio remains to be a challenging task[2,3,4,7].

In this paper, it has been tried to cope the above said problem. In this work it has been tried to achieve a better compression ratio and runs significantly faster than any existing compression program for Biological sequences. A lot of research work has already been carried out for developing programmes for Biological sequence compression. It is seen that all Biological sequence compression algorithms find repetition with in the sequence. Longer repetitive length implies higher compression gain. The compression ratio gain is high if highly similar subsequences are found. It is well known that there are similarities among different chromosome sequences. However, cross chromosomal similarities are seldom exploited in sequence compression[11]. The objective of this paper is to exploit self chromosomal similarity and cross chromosomal similarities as well. It should be noted that similar subsequences located within the chromosome sequence are called self similar while located in other chromosome sequence are called cross chromosome similar sequences.

2. METHODOLOGY

We use eight sequences to find chromosomal similarities. First, we search all cross similarities and then compress with the help of variable length LUT based compression algorithm. Large compression gain means that two subsequences are similar to each other.

2.1 Similarities Between Two Biological Sequences Chromosome

The potential gain in cross chromosomal compression is obtained by finding the total lengths of subsequence in the current chromosome sequence that is predicted from other chromosome sequence. The length of these cross reference subsequences determines the potential compression gain in multiple Biological sequence compression. Long length implies a high compression ratio.

2.2 Algorithm

Consider a finite sequence over the DNA alphabet {a, c, g, t}. Initialize an (n+1)x(m+1) array, L, for the boundary cases when i=0 or j=0. Namely, we initialize L[I,-1]=0 for i=-1,0,1,…,n-1 and L[-

1,j]=0 for j=-1,0,1,….,m-1. Then, we iteratively build up values in L until we have L[n-1,m-1], the length of a longest common subsequence of a finite sequence.

```
LCS(X,Y)
for  i ← -1 to n-1   do
          L[i,-1] ← 0
for  j ← 0 to m-1 do
          L[-1,j] ← 0
for  i ← 0 to n-1 do
          for j ← 0 to m-1 do
                    if X[i] = Y[j] then
                              L[i, j ] ← L[ i-1, j-1] + 1
                    else
                              L[i, j] ← max { L [i-1, j], L[i, j-1]}
return array L.
```

3. ENCODING REPEATS

Here, we use two step algorithms for compression: in first step we identify the all repetitions defined as Pre coding routine and in second step encoding algorithm is applied on both unique repeat and repeated subsequence.

Step 1: Pre Coding Routine

As discussed earlier the Pre coding routine help us to find the all repeats within a sequence, now method of pre coding routine algorithm will be presented.

i. Look Up Table(LUT)

The coding routine is based on variable length LUT, In which the initialization of table will be made first. We take all possible combination of three characters {a, t, g, or c} of the sequence which has been mapped onto a character chosen from the character set which consists of 8 bit ASCII characters. The generated LUT is given in table 1. It has been observed that with the help of above said generated table the implementation of pre coding routine becomes easy in handling the pre coding routine. Here it has been also observed that characters a, t, g, c and A, T, G, C will have same meaning. As an example if a segment “ACTGTCGATGCC” appeared in the LUT, in the destination file, it is represented as “j2X6”. By doing so the generated output will become case-sensitive.

It has been also observed that in Biological sequences some time a special character “N” appears. But it is exceptional. It will be necessary to consider this character also. Now handling this particular character will be dealt.

ii. Handling With the N

It has been realized that where ever the occurrence of N is repeated, there will be several Ns all together. To cater out this situation whenever the occurrence of Ns will take place in the sequences than the provision has been made in such a way that for its recognition a special character “/” inserted in the beginning and ending of Ns. All these Ns will be replaced by the characters representing total number of Ns. For example, if segment ”NNNNNN” will appear it will be replace by “/6/”.

We have taken three characters all together at a time in an input sequence. There is a possibility that while forming destination file by making a set of three characters all together, quite often less than three characters will be left in source file, for example, ATGATGATGCATTG and ATGATGATGCATTGC in which after making a set of three characters in first example TG and in second only character C is left over. So, how to handle such situation will be dealt now.

iii. Segment Which Consists of Less Than 3 Characters

In the Table 1 there is no appearance of `TC`. So to this situation just the original segment is written to destination file.

Step 2: Now, the development of algorithm will be described. There are seven steps in the designed algorithm. The steps from 1 to 6 do the task of approximate repeat and the last step signifies encoding. Here, w, wk and k represent character string,

Initialize: w=Nil.
Initialize: wk=Nil.
Initialize: k=Nil.

Step 1: read first three unprocessed characters (k). If k!= NULL, go along to step 2.Else (the EOF step 3 is reached),process the last one or two characters by step 5.

Step 2: Check that k has all non-N characters. If it is true, go to step 3.Else if k has N characters, go to step 4 Else go to step 5.

Step 3: If wk exists in the LUT then

```
        w=wk
Else
{
Output the code (character) for w
// ASCII code (character) that are mapped in LUT.
Add wk to the LUT table;
w=k;
}
End if;
```

Step 4: Search first N and successive Ns in the string and count total number of appearing in successive Ns, replace all the such Ns with "/n/" into destination file. After this, go to step 6. If number of successive Ns appears more than one time repeat the step 4.

Step 5: write non-N characters whose number is less than three into destination file directly without any modification. After that, go to step 6.

Step 6: Return to step 1 and repeat all process until EOF is reached.

Step 7: compress the output file by LZ77 algorithm.

Character	Base	Character	Base	Character	Base	Character	Base
! (33)	A A A	q (113)	T A A	' (39)	C A A	W(87)	G A A
b(98)	A A T	r (114)	T A T	H(72)	C A T	X(88)	G A T
" (34)	A A C	s (115)	T A C	I (73)	C A C	Y(89)	G A C
d(100)	A A G	$(36)	T A G	J (74)	C A G	Z(90)	G A G
e (101)	A T A	u (117)	T T A	K(75)	C T A	0 (48)	G T A
f (102)	A T T	v(118)	T T T	L(76)	C T T	1(49)	G T T
# (35)	A T C	w(119)	T T C	M(77)	C T C	2 (50)	G T C
h(104)	A T G	x(120)	T T G	N(78)	C T G	3(51)	G T G
i (105)	A C A	y (121)	T C A	O(79)	C C A	4 (52)	G C A
j (106)	A C T	z (122)	T C T	P(80)	C C T	5(53)	G C T
k (107)	A C C	%(37)	T C C	Q(81)	C C C	6 (54)	G C C
l (108)	A C G	B(66)	T C G	R(82)	C C G	7(55)	G C G
m(109)	A G A	&(38)	T G A	S (83)	C G A	8 (56)	G G A
n(110)	A G T	D(68)	T G T	((40)	C G T	9(57)	G G T
o (111)	A G C	E(69)	T G C	U(85)	C G C	+ (43)	G G C
p(112)	A G G	F(70)	T G G	V(86)	C G G	- (45)	G G G

TABLE 1: Initial Look up Table (LUT)

4. EXPERIMENTAL RESULT

Besides considering the total length of subsequences within a chromosome that can be referenced from other chromosomes, their distribution with in the sequence are also important. Let the subsequence in a sequence X that is similar to a subsequence in sequence i be X(i) and the subsequence j be X(j), the total length of subsequences within X that can be referenced from

i and j is given by T= | X(i) | + | X(j) | - | X(i) ∩ X(j) |. Obviously if these subsequences are well spread out such that | X(i) ∩ X(j) | is zero, i.e., they do not overlap in position, T maximized. This implies that a high proportion of the nucleotides within X can be predicted by cross referencing among chromosomes, resulting in a high compression gain.

The Biological sequences may be compressed by applying an algorithm based on fixed length look up table. But in this paper a different approach has been used to develop/ design the algorithm which is based on variable length look up table. Different Biological sequences have been taken as a sample and passed to developed program. The results obtained so have been given in tabular form as in Table 2. In the Table 2 the compressibility of some Biological sequences obtained by the method of algorithm based on variable length look up table has been also presented.

From Table-2 we can easily conclude the result that variable length length LUT compression method produces better compression, hence in the proposed algorithm we are using variable length LUT.

The analysis of Table-3 shows that the compression ratio is 91.87% which is the highest gain achieved by other existing algorithms.

Type of the Sequences	Original Size(bits) before compression	Size of the sequence after applying various compression algorithm			
		DNACompress	GenCompress	Fixed LUT	Variable LUT
Gallus β globin	752	272	360	256	248
Goat alanine β globin	732	256	352	248	232
Human β globin	752	272	360	256	248
Lemur β globin	760	280	376	264	256
Mouse β globin	776	280	376	264	256
Opossum β hemoglobin β- M gene	760	272	376	264	256
Rabbit β globin	736	264	352	256	248
Rat β globin	752	272	360	256	248
Avg	752.5	271	364	258	249

TABLE 2: Comparaisons between Different Biological Sequence Compression Techniques.

Type of Sequences	With reference to	Original Size(bits) before compression	Size of the sequence after applying compression algorithm	
			Cross chromosomal Similarity	Cross chromosomal Similarity Using variable length LUT
Gallus β globin	Human β globin	752	192	56
Goat alanine β globin	Rat β globin	732	192	56
Human β globin	Mouse β globin	752	176	56
	Rabbit β globin	752	176	56
Lemur β globin	Gallus β globin	760	40	16
	Opossum β hemoglobin β- M gene	760	56	16
Mouse β globin	Human β globin	776	208	72
Opossum β hemoglobin β- M gene	Goat alanine β globin	760	240	72
Rabbit β globin	Human β globin	736	136	48
Rat β globin	Mouse β globin	752	340	72
Avg		752.5	165.6	52

TABLE 3: Cross Chromosomal Similarity Using Variable length
Compression ratio=91.87%.
Bits/ Base=.5526 bits/base.

5. CONCLUSION AND DISCUSSION

There are several algorithms for the compression of biological sequence/genome. The widely used algorithms GenCompress, DNA Compress have the characteristics of simplicity & flexibility. Our proposed algorithm is also simpler and more flexible. All the existing algorithms are either statistics based or dictionary based. In this regard our algorithm is dictionary based. One more characteristic for our algorithm is that unlike other algorithm our algorithm is trying to compress whole genome structure.

The proposed algorithm is also very helpful to find the relatedness among different sequences. Also it is very useful in multiple sequence compression for both repetitive and non-repetitive. The result analysis of the application of our algorithm shows high compression ratio to other exiting Biological Sequence Compression. This algorithm also uses less memory compared to the other algorithm and its implementation is comparatively simple.

The proposed algorithm compresses all the Biological sequences having self chromosomal and cross chromosomal similarities. While all other algorithms only use one of the properties of sequences. If the sequence is compressed using proposed algorithm it will be easier to make sequence analysis between compressed sequences. It will also be easier to make multi sequence alignment. High compression ratio also suggests a highly similar sequences.

6. REFERENCES

1. Ateet Mehta , 2010, et al., *" DNA Compression using Hash Based Data Structure*", IJIT&KM, Vol2 No.2, pp. 383-386.

2. B.A., 2005, " *Genetics: A comceptual approach*." Freeman, PP 311.

3.Choi Ping Paula Wu, 2008, et al., " *Cross chromosomal similarity for DNA sequence compression*", Bioinformatics 2(9): 412-416.

4. Gregory Vey, 2009, "*Differential direct coding: a compression algorithm for nucleotide sequence data*", Database, doi: 10.1093/database/bap013.

5. J. Ziv and A., 1977, et al, "*A universal algorithm for sequential data compression*," IEEE Transactions on Information Theory, vol. IT-23.

6. K.N. Mishra, 2010, " *An efficient Horizontal and Vertical Method for Online DNA sequence Compression*", IJCA(0975-8887), Vol3, PP 39-45.

7. P. raja Rajeswari, 2010, et al., " *GENBIT Compress- Algorithm for repetitive and non repetitive DNA sequences*", JTAIT, PP 25-29.

8. Pavol Hanus, 2010, et al., "*Compression of whole Genome Alignments*", IEE Transactions of Information Theory, vol.56, No.2Doi: 10.1109/TIT.2009.2037052.

9. R. Curnow, 1989, et al. *"Statistical analysis of deoxyribonucleic acid sequence data-a review*," J Royal Statistical Soc., vol. 152, pp. 199-220.

10. Sheng Bao, 2005, et al. "A *DNA Sequence Compression Algorithm Based on LUT and LZ77*", IEEE International Symposium on Signal Processing and Information Technology.

11. U. Ghoshdastider, 2005, et al., "*GenomeCompress: A Novel Algorithm for DNA Compression*", ISSN 0973-6824.

12. Xin Chen, 2002, et al.," DNA *Compress: fast and effective DNA sequence Compression*" BIOINFORMATICS APPLICATIONS NOTE, Vol. 18 no. 12, Pages 1696–1698.

13. X. Chen, 2002, et al., "*Dnacompress:fast and effective dna sequence compression*," Bioinformatics, vol. 18,.

14. Voet & Voet, Biochemistry, 3rd Edition, 2004.

Face Alignment Using Active Shape Model And Support Vector Machine

Le Hoang Thai lhthai@fit.hcmus.edu.vn
Department of Computer Science
University of Science
Hochiminh City, 70000, VIETNAM

Vo Nhat Truong vntruong@gmail.com
Faculty/Department/Division
University of Science
Hochiminh City, 70000, VIETNAM

Abstract

The Active Shape Model (ASM) is one of the most popular local texture models for face alignment. It applies in many fields such as locating facial features in the image, face synthesis, etc. However, the experimental results show that the accuracy of the classical ASM for some applications is not high. This paper suggests some improvements on the classical ASM to increase the performance of the model in the application: face alignment. Four of our major improvements include: i) building a model combining Sobel filter and the 2-D profile in searching face in image; ii) applying Canny algorithm for the enhancement edge on image; iii) Support Vector Machine (SVM) is used to classify landmarks on face, in order to determine exactly location of these landmarks support for ASM; iv) automatically adjust 2-D profile in the multi-level model based on the size of the input image. The experimental results on Caltech face database and Technical University of Denmark database (imm_face) show that our proposed improvement leads to far better performance.

Keywords: Face Alignment, Active Shape Model, Principal Component Analysis.

1. INTRODUCTION

Face recognition is the problem to search human faces in large image database. In detail, a face recognition system with the input of an arbitrary image will search in database to output people's identification in the input image. The face recognition system's stages are illustrated in Figure 1 [5].

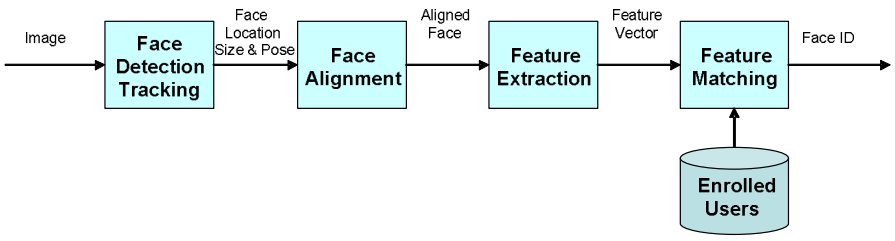

FIGURE 1: Structure of a face recognition system.

The face alignment is one of important stages of the face recognition. Moreover, face alignment is also applied for other face processing applications; such as face modeling and synthesis. Its objective is to localize the feature points on face images such as the contour points of eye, nose, mouth and face (illustrated in Figure 2).

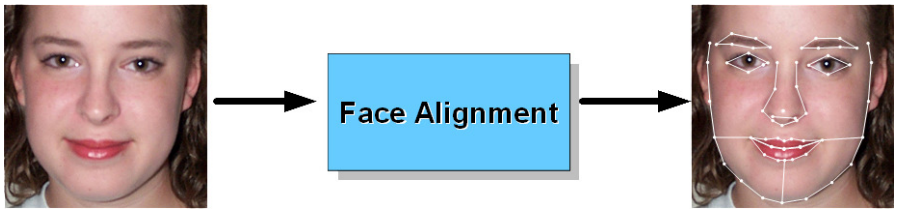

FIGURE 2: Face alignment.

There are many face alignment methods. Two popular face alignment methods are Active Shape Model (ASM) [16] and Active Appearance Model (AAM)[8] are proposed by Cootes et al.

The two methods use a statistical model to parameterize a face shape with Principal Component Analysis (PCA) method. However, their feature model and optimization are different. ASM algorithm has a 2-stage loop: in the first stage, given the initial labels, searching for a new position for every label point in its local region which best fits the corresponding local 1-D profile texture model; in the second stage, updating the shape parameters which best fits these new label positions. AAM method uses its global appearance model to directly conduct the optimization of shape parameters. Owing to the different optimization criteria, ASM performs more precisely on shape localization, and is quite more robust to illumination and bad initialization. In the paper extent, we develop the classical ASM method to create a new method named ASM-SVM which has achieved better results.

Because ASM only uses a 1-D profile texture feature, which is not enough to distinguish feature points from their local regions, the ASM algorithm often fall into local minima problem in the local searching stage. A few representative texture features and pattern recognition methods are proposed to reinforce the ASM local searching, e.g. Gabor wavelet, Haar wavelet, Ranking-Boost, Fisher Boost and MLP-ASM (Perceptron network) [5]. Nevertheless, an accurate local texture model to large data sets is still unachieved target.

In this paper, we propose the improvements in the local search of ASM. The main improvements are followed: first, build a model combining Sobel filter and the 2-D profile in searching face in image; second, applying Canny algorithms for the enhancement edge in image; third, support vector machine (SVM) is used to classify landmarks on face, in order to determine the exact location of these landmarks support for ASM; last, automatically adjust 2-D profile in the multi-level model based on the size of the input image.

The paper is structured as follows: Section 2, we present the classical ASM algorithm, section 3 presents details of our improvements and Section 4 presents experimental results and Section 5 presents conclusion and future works.

2. CLASSICAL ASM ALGORITHM

2.1 Training stage

A shape can be represented by n points $\{(x_i, y_i)\}$ as a 2n-D element vector, $X = (x_1, y_1, \ldots, x_n, y_n)^T$. With training shape S ($S = \{X_i\}$), we perform statistical shape on the same coordinates.

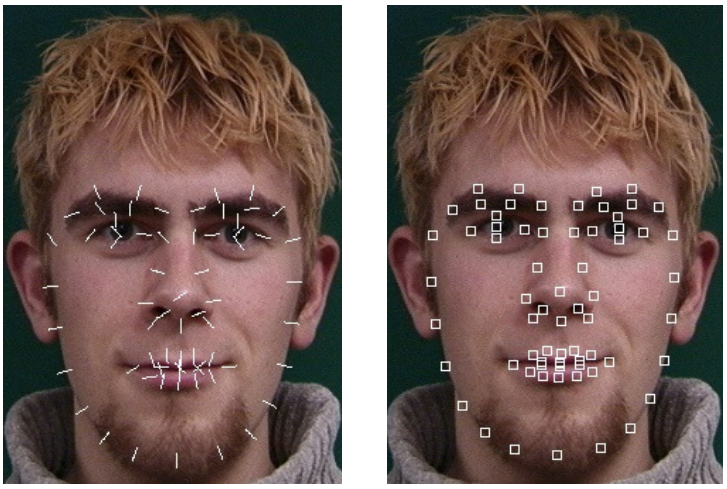

FIGURE 3: Local features to be built in the period of training (a) typical 1-D (b) typical 2-D

The shape of the training set S are aligned by algorithms Generalized Procrustes Analysis (GPA) [12]. Average shape ($\bar{x}$) is the average shape vector of all alignment shape. PCA is applied to calculate this shape and the covariance matrix is chosen so that accounted for 97.5% of the total value of training set that are arranged from large to small and used to store as the corresponding eigenvector matrix (P).

In next stage, we determine the gray level to create the statistical model of gray level around the landmarks to build a subspace represents the change of training shape. 1-D profiles are constructed by the gray level of points on the line with fixed length. These straight lines are orthogonal to the edge of this shape at the landmark. The gray level sample is stored as a vector that is then standardized by replacing each element of the vector with the intensity of gray levels (the difference of gray level at that point and the preceding point) and then dividing the magnitude of the vector average. Average profile (of all files in training set) is called average profile vector ($\bar{g}$) and covariance matrix of all the vector present as S_g. Average profile vector and covariance matrix are generated for each point and three level of the pyramid model (each image in the pyramid is half the image size of it before).

Similar, training data can be calculated by 2-D profile that is created at each landmark by the derivation of gray level image (the sum of square derivation in x and y directions) .Result matrix is transformed into a vector and then normalized by the sigmoid transformation for each specific element of profile g'_i as follow equation:

$$g'_i = \frac{g_i}{(|g_i| + q)} \quad (1)$$

q: const.

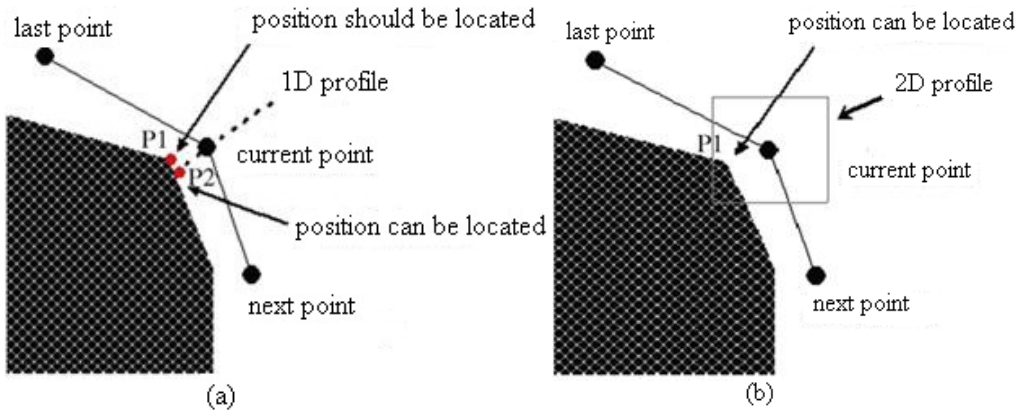

FIGURE 4: Illustrate local features: (a) typical 1D and (b) typical 2D

Using 1-D profile to find landmark in some cases is not accurate. For example, Figure 4(a) illustrates the case that desired position is P1. However, 1-D profile achieves P2 point instead of P1. Hence, 2-D profile is necessary to solve this problem (Figure 3 (b)). The desired location is P1 can be determined exactly by 2-D profile. Moreover, misplaced errors will reduce by using 2-D profile.

2.2 Alignment stage

In alignment phase, faces in test images will be identified in first step. Face detection algorithm, Viola Jones in OpenCV [11], is used for this step. After detecting the location of faces in images, similar transformations (scale, rotation and translation) will operate on the shape model that represent the face (constructing from the training data set) to fit this model to test face (the face is detected by OpenCV). Achieving shape will use as initial shape. A loop on the initial shape would be made to find suitable final landmark. These landmarks will form a shape that best suits the considered face image.

Typical multi-level model is built for image at each level by the method as in training phase. The process of identifying profile start from the lowest level of the pyramid (level 2) and gradually move up to the highest level (level 0) (Figure 4). Fluctuations of the landmarks are highest at the lowest level and they are smaller at higher levels. Best location of the landmark is determined by establishing the profile of the candidate neighboring around the landmarks. The candidate points that have nearly all features of average landmark will be selected as the new location of landmark. The weighting function that use in ASM to determine at this landmark is the smallest Mahalanobis distance ($f_1(g)$) of candidates (g) with average profile $\bar{g}$ by the following equation:

$$f_1(x) = (g - \bar{g})^T S_g^{-1} (g - \bar{g}) \quad (2)$$

Searching process on the 2-D profile with size 15x15 pixels around the landmark will operate at all levels of multi-level model.

When all landmarks move to the best location, a new shape (x_i) needs to be converted into an appropriate shape and to represent the boundary of face. This is done by Equation (3).

$$x_L = \bar{x} + Pb \quad (3)$$

x_L: the closest shape vector (x_I)
$\bar{x}$: average shape
P: eigenvector matrix

b: coefficient vector that is predicted to generate the face shape.

b is calculated by performing a loop so that the distance of formula (4) is the smallest.

$$dist(x_I, T(\bar{x} + Pb)) \quad (4)$$

T is a transformation that makes minimizing distance between x_I and $\bar{x} + Pb$. [9] represents the algorithm that finds b and T. When we get vector b, b_i is i^{th} element of vector b and it have to be between $-\alpha\sqrt{\lambda_i}$ and $+\alpha\sqrt{\lambda_i}$ with α is 3 and λ_i is i^{th} eigenvalue.

The limitation of these values can ensure that the generated shape is similar to those in the original training set.

At each level of multi-level model, a loop will be done until convergence (no significant change of landmark position in two consecutive loops). If convergence is done at lowest level,the shape will be changed by scale transformation and used as initial position for the next level of multi-level model. This process continues until convergence and achieving final landmarks at the highest level of multi-level model.

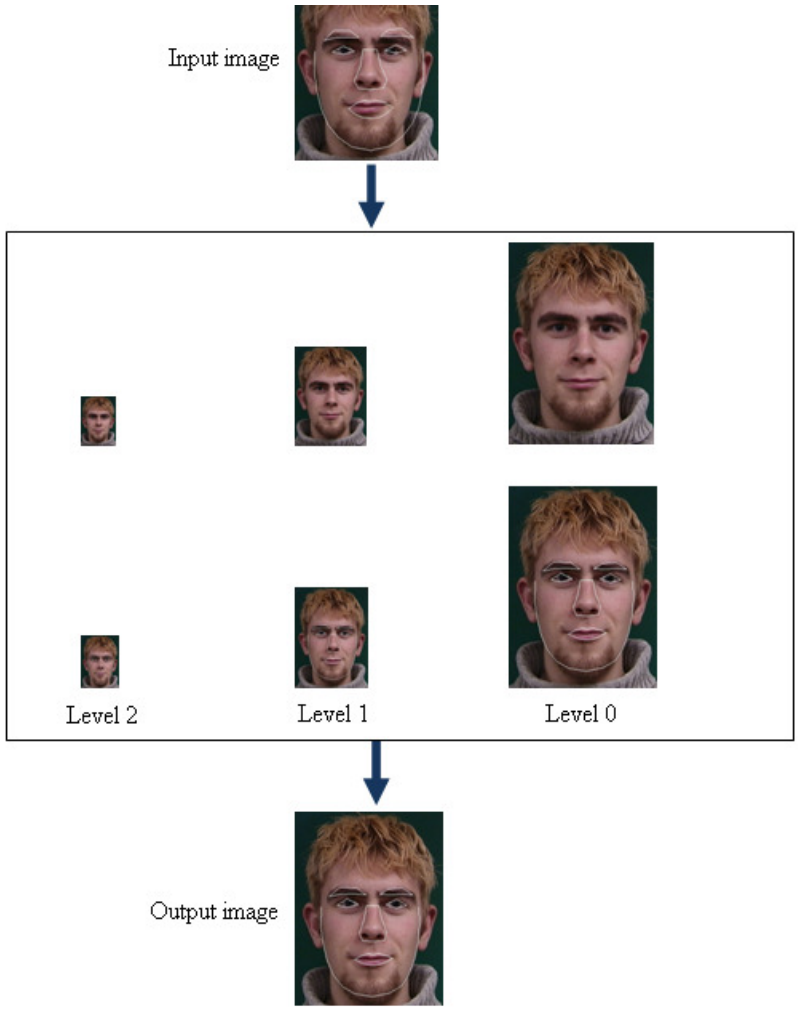

FIGURE 5: Illustrate alignment of multi-level model

3. IMPROVEMENT FOR ASM

3.1 Combining 2-D profile and Sobel filter

In order to balance brightness in image as well as distinguish between high and low frequency variations in image. In this paper, we determine the 2-D profile for each point by combining histogram equalization and Sobel filter as follow:

Step 1: Using the Histogram balancing algorithm to normalize brightness of image.
Step 2: Using the Sobel filter function in two directions x, y. Constructing texture matrix, with value of each point in the matrix is the square root of the sum of squared derivative in two directions (x and y).
Step 3: Normalizing result matrix to a vector by formula (5).

$$g'_i = \frac{g_i}{\sum g_i} \tag{5}$$

3.2 Enhancement edge by Canny algorithm

To increase more accurately for fitting points that lie along the boundary, we use the weighting function (6).

$$f_2(g) = (c - \mathrm{I})(g - \overline{g})^T S_g^{-1}(g - \overline{g}) \tag{6}$$

In the function $f_2(g)$, I is the gray value at candidate point and has value 0 (for the point not on the boundary) or 1 (for the point on the boundary). I determined based on enhancement edge by Canny algorithm [11]. c is a constant and we choose 2 for our experiments. Function $f_2(g)$ can increase the ability to searching landmark on the boundary of shape that hard to find in classical algorithm.

FIGURE 6: Illustrates the edge detection algorithm (Canny): (a) original image (b) resulting image

3.3 Applying SVM to find landmarks

In the classical ASM method, the PCA does not consider the distinction between the positive sample (the points represent the model (Section 4)) and negative sample (the points are not positive sample). So the searching landmark process often falls into local extreme values. To distinguish between positive sample and negative sample, we chose SVM method [10] because this method generalizes learning sample (without learning much data as other classification methods) and minimizes the structure error that increases the classified ability. In this paper, we use linear SVM.

For each point, we determine local 2-D profile with this. Next, the positive samples (the landmark) are selected from the focal point, whereas the negative samples (the point is not the landmark) randomly select window which has same size and different focal point with positive sample.
Algorithms search candidate around the current landmark to determine the new landmark:

Input: shape X $\{(x_i, y_i)\}$
Output: shape X'$\{(x'_i, y'_i)\}$
For each point (x_i, y_i) of X
For each window has focus point (x', y') that belong to window with focus point (x_i, y_i).
Applying SVM in the window (x', y'). If the return value is +1, the point (x', y') lies on the boundary, otherwise returns -1.
Selecting a point (x ', y') that value of function $f_2(g)$ is the smallest. This point is the new landmark.

3.4 Adjusting profile length

In the classical ASM algorithm, the lengths of profile windows are the same size at each level in multi-level model. However, from experiment, we found that the shift of the points in each level is different. At higher levels, the shift is smaller. Moreover, the shift is very small when the candidates are close to destination. From this observation, we adjust profile length according to different levels. Adjusting profile length save the computational cost and increase the accuracy of landmark determination. Length of the window is our proposal to reduce the level of ascent. The length of the first level is L, L/2 for level 2, and L/4 for the final level. In this experiment we use a length for the first level is 15 pixels.

4. EXPERIMENTAL RESULTS

Face shape are made from 68 landmarks that extract with specific groups as follow: face boundary (15 points), right eyebrow (8 points), left eyebrow (8 points), left eye (8 points), right eye (8 points), nose (9 points) and oral (12 points).

Evaluating performance of our proposed method and other methods, we use the average error calculation function as follow:

$$E_{ave} = \frac{1}{n} \times \frac{1}{k} \sum_{i=1}^{n} \sum_{j=1}^{k} abs(x(i, j) - pos(i, j)) \quad (7)$$

4.1 CalTech database

Caltech face image data includes 450 jpeg images with 896 x 592 sizes of 27 subjects.
We randomly select 300 images for training and the remaining 150 images for test. Table 1 illustrates the test results.

Method	Average Error (E_{ave})
ASM	14.021
MLP-ASM	12.403
ASM-SVM	10.548

TABLE 1: Experimental results on the Caltech database

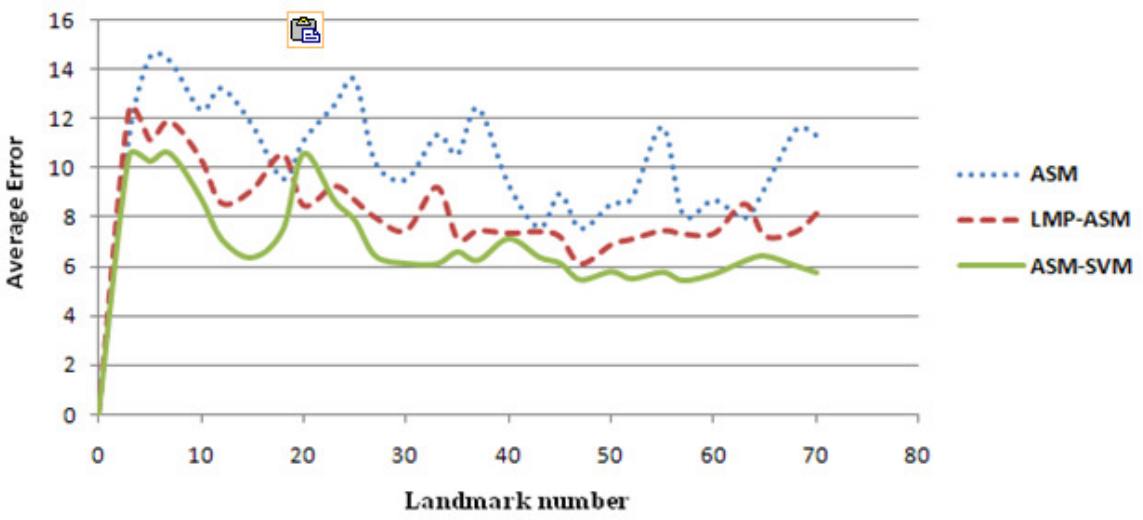

FIGURE 7: Comparison between the classical ASM, MLP-ASM and ASM-SVM.

FIGURE 8: Some experimental results

4.2 DTU database

DTU face image data includes 240 jpeg images with 640 x 480 sizes of 40 subjects.

We randomly select 160 images for training and the remaining 80 images for test. Table 2 illustrates the test results.

Method	Average Error (E_{ave})
ASM	11.751
MLP-ASM	9.147
ASM-SVM	7.176

TABLE 2: Experimental results on the DTU database

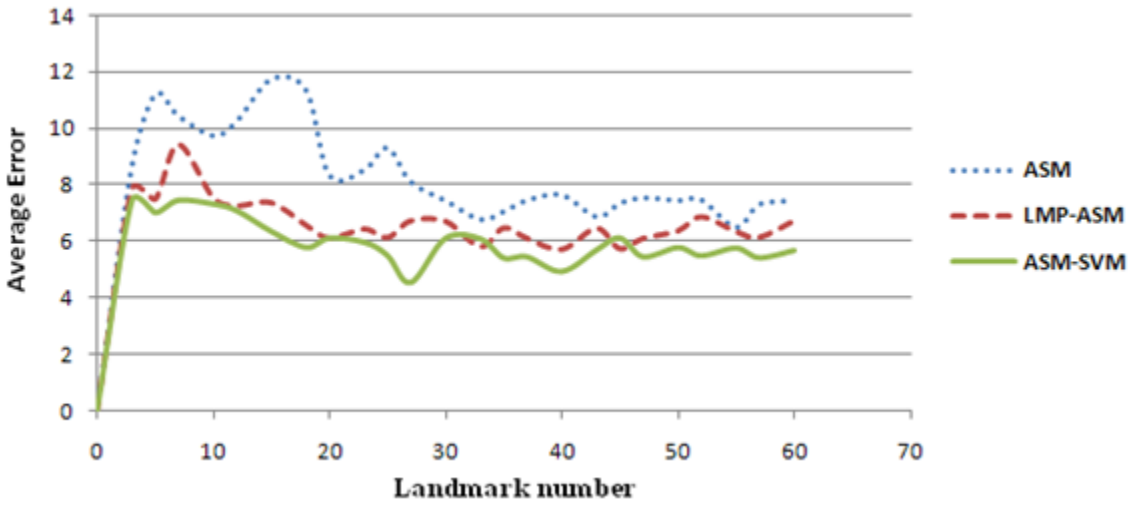

FIGURE 9: Comparison between the classical ASM, MLP-ASM and ASM-SVM.

FIGURE 10: Some experimental results

4.3 Others

FIGURE 11: Results on real images obtained from the internet.

Reviews: with experimental results on two databases: Caltech and DTU, our approach is general and can be applied to many different databases. With using edge detection methods, our method gets high efficiency when compare to classical ASM at landmarks on the boundary of face.

5. CONSLUSION & FUTURE WORKS

In this paper, we propose an alignment model using ASM combine with SVM. Instead of using 1-D profile model, we use 2-D profile model and combine with Sobel filter function to new landmarks that are neighbored with original ones. This model is useful for finding landmarks, which bases on the strong classifier of SVM and the distance measuring of Mahalanobis, as well as determine strong edges to increase the accuracy of determining landmarks. General and powerful classifier of proposed model makes ASM more efficient. The result of the comparison proposed method to classical ASM, bases on Caltech database and DTU database (imm_face), show that our proposed improvements are better performance.

In the future, we can use hierarchical approach. Firstly, using ASM for the global features (face boundary), and then we use ASM for the local features (left eye, right eye, nose, mouth). Combining this method with Expectation Maximization Algorithm is useful for adjusting incorrect landmarks.

6. REFERENCES

1. Kwok-Wai Wan, Kin-Man Lam, Kit-Chong Ng. “An accurate active shape model for facial feature extraction”. Journal of Pattern Recognition Letters, Volume 26, Issue 15, November 2005.

2. Zhonglong Zheng , Jia Jiong, Duanmu Chunjiang, XinHong Liu, Jie Yang. “Facial feature localization based on an improved active shape model”. International Journal of Information Sciences, Volume 178, Issue 9, May 2008.

3. Chunhua Du, Qiang Wu, Jie Yang, Zheng Wu. “SVM based ASM for facial landmarks location”. IEEE 8th International Conference on Computer and Information Technology, pp. 321-326, 2008.

4. G.J. Edwards, T.F. Cootes, and C.J. Taylor. “Face Recognition Using Active Appearance Models”. Proceedings of the 5th European Conference on Computer Vision-Volume II ,1998.

5. Le Hoang Thai, Bui Tien Len. “Local texture classifiers based on multi layer perceptron for face alignment”. Information and Communication Technology in Faculty of Information Technology 2008, ICTFIT 2008, pp104-113, 2008.

6. Liting Wang, Xiaoqing Ding, Chi Fang. “Generic face alignment using an improved active shape model”. International Conference on Audio, Language and Image Processing, ICALIP 2008, pp. 317-321, 2008.

7. Stephen Milborrow and Fred Nicolls. “Locating facial features with an extended active shape model”. Proceedings of the 10th European Conference on Computer Vision: Part IV, ECCV '08, 2008.

8. Timothy F. Cootes, Gareth J. Edwards, and Christopher J. Taylor. “Active Appearance Models”. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 23, no. 6, June 2001.

9. T. F. Cootes and C. J. Taylor. “Statistical Models of Appearance for Computer Vision”. Technical Report, Imaging Science and Biomedical Engineering, University of Manchester, March 2004.

10. C. Cortes and V. Vanpik. “Support Vector Networks”. Machine Learning, pp. 273-297, 1995.

11. Intel: Open Source Computer Vision Library, Intel, 2007.

12. J. C. Gower.“Generalized Procrustes Analysis”. Psychometrika, vol. 40, no. 1, pp. 33-51, March 1975.

13. Ralph Gross, Iain Matthews, Simon Baker. “Active appearance models with occlusion”. Image and Vision Computing, Volume 24, Issue 6, pp. 593-604, 2006.

14. Shuicheng Yan, Ce Liu, Stan Z. Li, Hongjiang Zhang, Heung-Yeung Shum, Qiansheng Cheng. “Face alignment using texture-constrained active shape models”. Image and Vision Computing, Volume 21, Issue 1, pp. 69-75, 2003.

15. Stan Z. Li, Anil K. Jain. “Handbook of face recognition”. Springer Science and Business Media Inc, 2005.

16. Tim Cootes. “An Introduction to Active Shape Models”. Chapter 7, Image Processing and Analysis, pp.223-248, Oxford University Press, 2000.

17. Xinbo Gao, YaSu, XuelongLi, DachengTao. “Gabor texture in active appearance models”. Neurocomputing, Volume 72, Issues 13-15, pp. 3174-3181, August 2009.

CALL FOR PAPERS

Journal: International Journal of Biometrics and Bioinformatics (IJBB)
Volume: 5 **Issue:** 1
ISSN:1985-2347
URL: http://www.cscjournals.org/csc/description.php?JCode=IJBB

About IJBB

The International Journal of Biometric and Bioinformatics (IJBB) brings together both of these aspects of biology and creates a platform for exploration and progress of these, relatively new disciplines by facilitating the exchange of information in the fields of computational molecular biology and post-genome bioinformatics and the role of statistics and mathematics in the biological sciences. Bioinformatics and Biometrics are expected to have a substantial impact on the scientific, engineering and economic development of the world. Together they are a comprehensive application of mathematics, statistics, science and computer science with an aim to understand living systems.

We invite specialists, researchers and scientists from the fields of biology, computer science, mathematics, statistics, physics and such related sciences to share their understanding and contributions towards scientific applications that set scientific or policy objectives, motivate method development and demonstrate the operation of new methods in the fields of Biometrics and Bioinformatics.

To build its International reputation, we are disseminating the publication information through Google Books, Google Scholar, Directory of Open Access Journals (DOAJ), Open J Gate, ScientificCommons, Docstoc and many more. Our International Editors are working on establishing ISI listing and a good impact factor for IJBB.

IJBB List of Topics

The realm of International Journal of Biometrics and Bioinformatics (IJBB) extends, but not limited, to the following:

- Bio-grid
- Bioinformatic databases
- Biomedical image processing (registration)
- Biomedical modelling and computer simulation
- Computational intelligence
- Computational structural biology
- Bio-ontology and data mining
- Biomedical image processing (fusion)
- Biomedical image processing (segmentation)
- Computational genomics
- Computational proteomics
- Data visualisation

- DNA assembly, clustering, and mapping
- Fuzzy logic
- Gene identification and annotation
- Hidden Markov models
- Molecular evolution and phylogeny
- Molecular sequence analysis
- E-health
- Gene expression and microarrays
- Genetic algorithms
- High performance computing
- Molecular modelling and simulation
- Neural networks

IMPORTANT DATES

Volume: 5
Issue: 1
Paper Submission: January 31, 2011
Author Notification: March 01, 2011
Issue Publication: March /April 2011

CALL FOR EDITORS/REVIEWERS

CSC Journals is in process of appointing Editorial Board Members for ***International Journal of Biometrics and Bioinformatics***. CSC Journals would like to invite interested candidates to join **IJBB** network of professionals/researchers for the positions of Editor-in-Chief, Associate Editor-in-Chief, Editorial Board Members and Reviewers.

The invitation encourages interested professionals to contribute into CSC research network by joining as a part of editorial board members and reviewers for scientific peer-reviewed journals. All journals use an online, electronic submission process. The Editor is responsible for the timely and substantive output of the journal, including the solicitation of manuscripts, supervision of the peer review process and the final selection of articles for publication. Responsibilities also include implementing the journal's editorial policies, maintaining high professional standards for published content, ensuring the integrity of the journal, guiding manuscripts through the review process, overseeing revisions, and planning special issues along with the editorial team.

A complete list of journals can be found at http://www.cscjournals.org/csc/byjournal.php. Interested candidates may apply for the following positions through http://www.cscjournals.org/csc/login.php.

Please remember that it is through the effort of volunteers such as yourself that CSC Journals continues to grow and flourish. Your help with reviewing the issues written by prospective authors would be very much appreciated.

Feel free to contact us at coordinator@cscjournals.org if you have any queries.

Contact Information

Computer Science Journals Sdn BhD
M-3-19, Plaza Damas Sri Hartamas
50480, Kuala Lumpur MALAYSIA

Phone: +603 6207 1607
+603 2782 6991
Fax: +603 6207 1697

BRANCH OFFICE 1
Suite 5.04 Level 5, 365 Little Collins Street,
MELBOURNE 3000, Victoria, AUSTRALIA

Fax: +613 8677 1132

BRANCH OFFICE 2
Office no. 8, Saad Arcad, DHA Main Bulevard
Lahore, PAKISTAN

EMAIL SUPPORT
Head CSC Press: coordinator@cscjournals.org
CSC Press: cscpress@cscjournals.org
Info: info@cscjournals.org

COMPUTER SCIENCE JOURNALS SDN BHD
M-3-19, PLAZA DAMAS
SRI HARTAMAS
50480, KUALA LUMPUR
MALAYSIA