INTERNATIONAL JOURNAL OF COMPUTER APPLICATIONS IN ENGINEERING, TECHNOLOGY AND SCIENCES (IJ-CA-ETS)

CAPITALIZING DATA MINING IN BIOINFORMATICS FOR BIOLOGICAL DATA ANALYSIS

[1] MS. RICHA MEHTA, [2] MR. RINAL DOSHI, [3] DR. AASHISH N. JANI

[1] Research Scholar, Singhania University, Rajasthan, Lecturer SKPIMCS-MCA Department, KSV University, Gandhinagar, Gujarat, India
[2] Research Scholar, School of Engineering, R.K.University, Rajkot, Lecturer L.C.Institute of Technology, GTU, Bhandu, Gujarat, India
[3] Research Guide, Asst. Prof. SKPIMCS-MCA Department, KSV University, Gandhinagar, Gujarat, India

triumphricha@gmail.com, rinal.doshi@lcit.org,aashishjani@yahoo.com

ABSTRACT - Data mining is an emerged field. Conventional data analysis techniques offer limitations to process efficiently very large volume of data. The effective data analysis technique to manage and process large volume of data is Data Mining which is ultimately a Knowledge Discovery technique. These techniques are usually coupled with databases to carry out Data Analysis efficiently, it because of the above fact the data is well organized and gets the support of efficient access and storage of intermediate results for later access. Data Mining shares its roots in Databases, Statistics, Machine Learning, Pattern Recognition, Image Processing etc. Data Mining has an emerged interdisciplinary field of research named as Bioinformatics where Biological Databases are coupled with Data Mining Techniques. All these interdisciplinary fields have experienced over a period of time faster adoption and implementation of Emerging Technologies that have made possible the setting up of experiments at large scale and these experiments are resulting in huge growth in the amount of Biological Data. This increasing growth of the biological data is demanding high performance computers systems for the purpose of analysis and visualization of data for deeper understanding. In this paper the attempt is made to discuss the use of Biological Data mining techniques, tools and challenges to analyse voluminous biological data related to Structural Bioinformatics, Bioinformatics Text mining, Biological sequence Analysis and Gene Expression Analysis.

KEYWORDS – Data Mining, Knowledge Discovery Technique, Bioinformatics, Biological Databases, Medicalinformatics, Emerging Technologies, Biological Data mining techniques, tools and challenges, Structural Bioinformatics, Bioinformatics Text mining, Biological sequence Analysis and Gene Expression Analysis.

I. INTRODUCTION

Bioinformatics is field of science in which biology, computer science and Information Technology are merged to form single discipline. The converged discipline of Bioinformatics has lead the researcher for the discovery of hidden biological insights from the large biological data sets. The beginning of genomic revolutions was limited to creation and maintenance of databases to store biological information. Today these databases have taken the shape of very large databases and the development of tools that help to analyse these data in detail to gain much deeper biological insights.

Luscombe et al. [1] identified the aims of Bioinformatics as follows:

- The design of the organization of data structure that supports elegant access of stored information to the researchers and facilitates the submission of newly generated data pertaining to the respective field.
- The development of tools that help in the analysis of data.
- The use of these tools to analyse the individual systems in detail, in order to gain new biological insights.

There is a strong research interest in methods of data mining with knowledge discovery to generate models of biological systems. There is a need to develop scalable as well as more efficient data mining systems that are expected to contribute for the enhancement to our present understanding of biological systems. This will lead to build up knowledge a discovery system that is expected to help us to carry out biological research more vigorously

II. BIOLOGICAL DATA MINING TECHNIQUES AND TOOLS

Although Biological Data Mining can be considered under "application exploration" or "mining complex

types of data," the unique combination of complexity, richness, size, and importance of biological data warrants special attention in data mining.

III.I DATA MINING IN STRUCTURAL BIOINFORMATICS

For the Analysis and prediction of three dimensional structures of Biological macromolecules like proteins, there is a subfield of Bioinformatics named as Structural Bioinformatics. As structural data are not linear, doing data mining in structural Bioinformatics is a herculean task. Moreover, the search space for most structural problems is continuous, namely infinite and demands highly efficient and heuristic algorithms.

Various protein properties such as active sites, modification sites, localization, stability, globularity, shape, protein domains, secondary structure and interactions are being predicted with the use of different machine learning and data mining algorithms [4].

Neural networks, nearest neighbour classifiers and hierarchical clustering algorithms are the most famous methods for this process.

Protein secondary structure prediction uses Machine learning and data mining methods. Over the past 35 years this problem is being studied and so many different techniques have created developed. Initially, statistical approaches were adopted to deal with this problem. Later, more accurate techniques based on information theory, Bayes theory, nearest neighbor, and neural networks were developed. Combined methods such as integrated multiple sequence alignments with neural network or nearest neighbor approaches improve prediction accuracy.

Other important problems of structural Bioinformatics that utilize machine learning and data mining methods are the RNA secondary structure prediction, the inference of a protein's function from its structure, the identification of protein-protein interactions and the efficient design of drugs, based on structural knowledge of their target.

Some of the data mining tools for structural Bioinformatics are as follows:

- PIG: It is a 3D protein structural web service. It includes Geno3D, Modeome3D, MAGOS and SuMo. Geno3D is used for automatic molecular modeling of protein 3D structures. Modeome3D is an interactive system for the integrated analysis of phenotypic consequences of missense mutations in proteins involved in human genetic diseases. MAGOS is for automated protein modelling coupled to the creation of a hierarchical and annotated Multiple Alignment of Complete Sequences
- KAKSI : Used for Assignment of protein secondary structures
- GORV and OSS-HMM: Used for prediction of protein secondary structures
- VAST : Used for 3D structure comparisons

III.II BIOINFORMATICS TEXT MINING

Information science, Bioinformatics and Computational Linguistics fields have come up with the new hybrid discipline that have evolved from the Automatic extraction of information about genes, proteins and their functional relationships from text documents, which is main potential for the scope of Text mining in molecular biology. In the field of research it is important for researchers to gain access to the appropriate and clean information. Latest technologies for information retrieval, Symentic web and text mining are used in current research practise. Ontology, an interaction and function between biological entities is textome, is the new term for text body for extraction of information used by Bioinformatics experts [5].

Several studies have categorized the tasks of BTM from different points of view. Cohen and Hersh [6] provide a high-level categorization, identifying the main tasks to be the following:

- Named Entity Recognition (NER).
- Text classification.
- Synonym and abbreviation extraction.
- Relationship extraction.
- Hypothesis generation.

Some of the data mining tools for Bioinformatics Text Mining are as follows:

- KLEIO - an advanced information retrieval system providing knowledge enriched searching for biomedicine.
- U-Compare - U-Compare is an integrated text mining/natural language processing system based on the UIMA Framework, with an emphasis on components for biomedical text mining.
- MEDIE - an intelligent search engine to retrieve biomedical correlations from MEDLINE, based on indexing by Natural Language Processing and Text Mining techniques
- GENIA tagger - Analyses biomedical text and outputs base forms, part-of-speech tags, chunk tags, and named entity tags
- NEMine - Recognises gene/protein names in text

III.III BIOLOGICAL DATA ANALYSIS AND MINING

The huge volume of Biological data requires High performance computing platforms, software tools on this platform to carry out Data mining, Data Analytics and Visualization.

III.III.I BIOLOGICAL SEQUENCE ANALYSIS

As large number of sequencing projects has been completed there is a need to annotate those genomes [4]. It is observed that the paradigm shift is taking place from static structural genomics to dynamic functional genomics. Importance is given to the functional informational assignment to known sequences. Gene prediction is targeted to identification of DNA stretches that are biological function. The Gene prediction in case of Eukaryotic

genomes is a complex task.

Eukaryotic gene consists of Exons which are coding paths. These exons when separated are called introns. The separation is by way of intervening non coding sequences. Introns are removed from transcribed RNA by process of RNA splicing. The challenging task is the recombination of splice sites which are boundaries between adjacent exons and introns. Still more challenging task appears in case of alternative splicing possibilities. This alternative splicing demands the application of predictive data mining methodologies to get solution of such complex problem.

Another reason that gene prediction is easier in prokaryotic genomes is the presence of known conserved patterns around various signals, like promoters, transcription start sites, and translation initiation sites of prokaryotic sequences.

The most important sequence analysis tasks that exploit machine learning and data mining algorithms are the following:

- Prediction of regulatory regions (i.e. promoters and enhancers), which are segments of DNA where regulatory proteins bind preferentially and thus control gene expression and consequently protein expression.
- Prediction of the transcription start site, where the process of transcription starts.
- Prediction of the translation initiation site, where the process of translation initiates.
- Prediction of the splice sites for determining exons and introns
- Prediction of polyadenylation sites where a polyA (multiple adenines) tail is added at the mRNA sequence. Alternative polyadenylation makes the problem more challenging.
- Prediction of coding region in Expressed Sequence Tags (ESTs) that are short and partial sequences of transcribed spliced mRNAs.
- Comparison of a sequence with a database of known sequences for finding possible homologies (e.g. close evolutionary relationships) and grouping of structurally related sequences.

Many data mining techniques have been utilized to deal with the above problems. Usually most of these techniques have to be modified and adapted so that can be efficiently and effectively applied to this kind of problems. The most common include neural networks, Bayesian classifiers [7], decision trees, and support vector machines [8], as well as multiple classifier systems [9].

Some of the data mining tools for Biological Sequence Analysis are as follows:

- BLAST: local search with fast k-tuple heuristic (Basic Local Alignment Search Tool)
- FASTA: local search with fast k-tuple heuristic, slower but more sensitive than BLAST
- HMMER: local and global search with profile Hidden Markov models,
- HHpred / HHsearch: pairwise comparison of profile Hidden Markov models; very sensitive, but can only search alignment databases

III.III.II GENE EXPRESSION ANALYSIS

A Number of genes are contained in an organism. These genes code the synthesis of protein molecules or mRNA. Usually every cell in an organism has the same set of genes and chromosomes. However there are few exceptions to it. The two cells may have different property and functions due to the difference in protein abundance. The abundance of a protein is partly determined by the levels of mRNA which in turn are determined by the expression levels of the corresponding gene. The process of conversion of the information encoded in a gene, first into mRNA and then to a protein structures and functions of a cell is called gene expression. A tool for analyzing gene expression is microarray or gene chip. A microarray experiment [10]measures the relative mRNA levels of typically thousands of genes, providing the ability to compare the expression levels of different biological samples. These samples may correlate with different time points taken during a biological process or with different tissue types such as normal cells and cancer cells [11]. Another method for measuring gene expression is Serial Analysis of Gene Expression (SAGE) [12], which allows the quantitative profiling of a large number of mRNA transcripts.

The greatest challenge posed by gene expression data is that they contain a small number of samples (less than a hundred), and a very large number of features (genes), that is typically in thousands. Many feature selection approaches have been utilized for reducing the dimensionality of the data by selecting a small number of genes [13]. Moreover, a large number of genes are usually irrelevant and uninformative for the classification. The danger of overshadowing the contribution of relevant genes is reduced when gene selection is applied.

Clustering is the far most used method in gene expression analysis. Clustering methods can be used to cluster genes with similar behaviour or samples with similar gene expressions together. A category of clustering algorithms, called subspace clustering or bi-clustering algorithms, are used to simultaneously cluster genes and samples. Hierarchical clustering is another frequently applied method in gene expression analysis. An important issue concerning the application of clustering methods in microarray data is the assessment of cluster quality[10]. Many techniques such as bootstrap, repeated measurements, mixture model- based approaches, sub-sampling and others have been proposed to deal with the cluster reliability assessment. External criteria have also been used for validating clusters quality based on predefined partitions of the data [14].

Some of the data mining tools for Gene Expression Analysis are as follow:

- GEM-TREND - retrieve gene expression

data by comparing query gene-expression pattern with those of GEO gene expression data

- J-Express - Portable software package for multidimensional scaling, clustering, and visualization of microarray data .
- GSCope -SOM custering and Gene-Ontology Analysis of microarray data
- ScanAlyze, Cluster, TreeView - Gene analysis software from the Eisen Lab
- GeneLinker - Gene expression and proteomics analysis software.
- GeneSpring - Gene expression analysis software from Silicon Genetics
- GEMS (Gene Expression Mining Server) simple but promising new approach for biclustering based on a Gibbs sampling paradigm.
- GEPAS (Gene Expression Pattern Analysis Suite) - an experiment-oriented pipeline for the analysis of microarray gene expression data. It contains an incredible number of tools for normalization, preprocessing, viewing, clustering, differential expression, supervised classification, and data mining & analysis.

IV BIOLOGICAL DATA MINING AND ANALYSIS

The procedure is presented as the experimental work in the Scientific Computing Software Tool Environment to carry out Analytics for Gene Expression Profiles.

A sample dataset is considered for visualizing the patterns in gene expression profiles, using gene expression data from yeast shifting from fermentation to respiration. The sample data set is obtained from the data source: The Gene Expression Omnibus Series GSE28. The implementation of gene expression analysis on a dataset can be done according to the below procedure.

IV.I DATA SET EXTRACTION

The process imports the data from the Data source-Gene Expression Omnibus Series GSE28. The data is transformed into the format acceptable to the software tool analysing the data.The extracted data file contains outcomes of the experiment performed stepwise with seven times repetition, Gene Names as well as corresponding array of times at which expression levels were recorded.

The extracted data set is loaded into the tool environment:

```
>> load rm_yeastdata.mat
```

The size of the data is obtained with the use of the function:

```
>> namel(genes)
```

Displays size 6400

The cell array elements can be accessed using cell array indexing

```
>> genes{15}
YAL054C
```

Displays the 15th row of the variable yeastvalues, which contains expression levels for the open reading frame YAL054C.

IV.II FILTERING GENES

The process is targeted to filter the data by the removal of genes which are not expressed or does not change in order reduce the number of expression profiles to subset that contains interesting most significant genes.

On inspecting gene list it may be observed that there are some empty spots on array marked as 'EMPTY'. These points can be located using function strcmp and can be eliminated.

```
>>rm_emptySpots=…
strcmp('EMPTY',genes);
>>yeastvalues(rm_emptySpots…      :) = [];
>>genes(rm_emptySpots) = [];
>> numel(genes)
```

Displays reduced size 6314

The yeastvalues data may contain at several places expression level marked as NaN (Not a Number), which indicates that no data was collected for that spot at that time step. Such missing values are taken care by replacing them with the mean or the median of the data for that particular gene.

The isnan function is used to identify the genes with missing data and to remove the genes isnan commands is used.

```
>>rm_nanIndices=...    any(isnan(yeastvalues),2);
>>yeastvalues(rm_nanIndices… :) = [];
>>genes(rm_nanIndices) = [];
>> namel(genes)
```

Displays reduced size 6276

To filter the genes over the time with small variance genevarfilter function is used which returns a variable gene similar sized logical array having 1's denoting row of yeastvalues showing variance more than 10th percentile and 0's having below the threshold.

```
>>rm_mask=…
genevarfilter(yeastvalues);
%Use the rm_mask as an index %into the values to
remove %the filtered genes.
>>rm_yeastvalues=…
yeastvalues(rm_mask,:);
>> genes = genes(rm_mask);
>> namel(genes)
```

Displays reduced size 5648

The very low absolute expression valued genes are further removed by genelowvalfilter function which also calculated the filtered data and names.

```
>>[rm_mask,yeastvaluesgenes…
]=Genelowvalfilter…
(yeastvalues,genes,…
'absval,log2(4));
>> numel(genes)
```

Displays reduced size 423

The low entropy genes profile are removed further using geneentropyfilter function:

```
>>[rm_mask,yeastvalues,…
```

```
genes]=Geneentropyfilter…
(yeastvalues,genes,…
'prctile, 15 );
>> numel(genes)
```

Displays reduced size 310

IV.III CLUSTERING GENES

The way we get a manageable list of genes. Now we can look for relationships between the profiles utilizing clustering techniques of the Statistics Toolbox.To achieve hierarchical clustering pdist function is used to compute the pairwise distances between the profiles, and the function linkage facilitates to create the hierarchical cluster tree.

```
>>rm_corrDist=…
pdist(yeastvalues,…    'corr');
>>rm_clusterTree=linkage…
(rm_corrDist,'average');
```

The cluster function computes the clusters based on cutoff distance or maximum number of clusters. The following use of the function with 'maxclust' option to identify 16 distinct clusters.

```
>>rm_clusters=cluster…
(rm_clusterTree,'maxclust'…
16);
```

The above cluster profiles of the genes can be plotted together with the use of simple loop and the function script using subplot function.

```
figure
for c = 1:16
subplot(4,4,c);
plot(times,yeastvalues((clusters==…    c),:)');
axis tight
end
```

The MATLAB Environment displays the underneath subplots on execution of the script

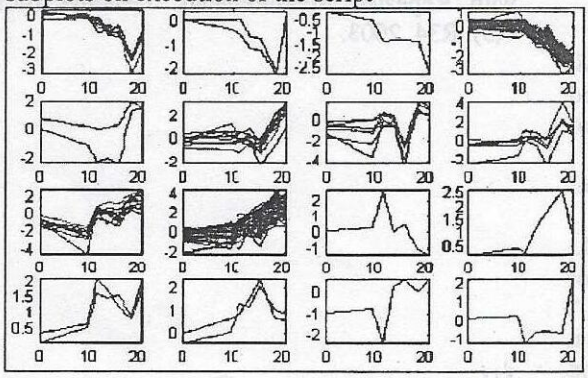

Figure 1: Hierarchical Clustering of Profile

The K-Means clustering function supported in Statistics Toolbox is used to create 16 clusters. The K-means clustering algorithm is different Hierarchal clustering and generates clusters not necessarily the same clusters as obtained in hierarchical clustering. The following script generates K-Means clusters.

```
[cidx,ctrs]=kmeans…
(yeastvalues,16,'dist',…
'corr','rep',5,'disp'…
'final');
figure
for c = 1:16
subplot(4,4,c);
plot(times,…
yeastvalues ((cidx==…
c),:)');
axis tight
end
```

The MATLAB Environment displays the underneath subplots on execution of the script.

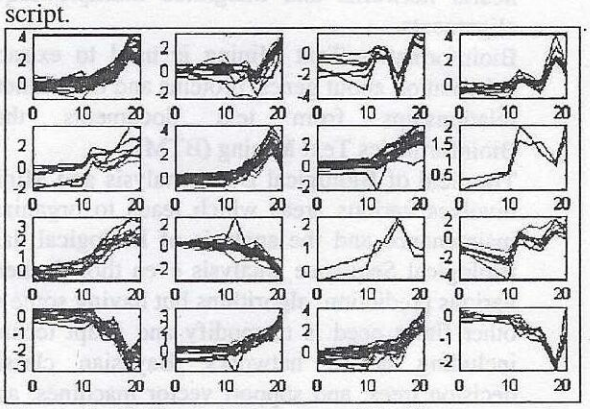

Figure 2: K-Means Clustering of Profile

If on needs to display only the centroids the following script serves the purpose .

```
figure
for c = 1:16
subplot(4,4,c);
plot(times,ctrs(c,:)';
axis tight
axis off
end
```

The MATLAB Environment displays the underneath subplots on execution of the script.

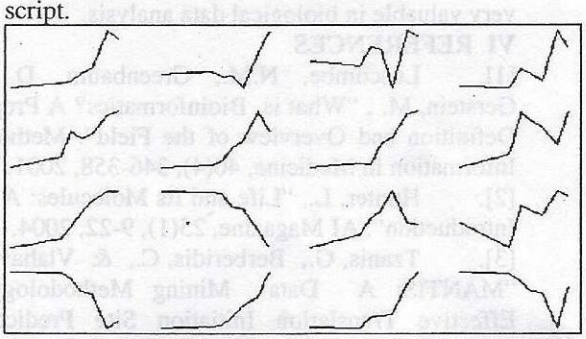

Figure 3: K-Means Clustering of Profile displaying centroids

V CONCLUSION

The recent technological advances, have led to an exponential growth of biological data. Scientists often have to use exploratory methods instead of confirming already suspected hypotheses. Data Mining aims to provide the analysts with novel, effective and efficient computational tools to overcome the obstacles and constraints posed by the traditional Statistical Methods. Feature selection, normalization of the data, visualization of the results and evaluation of the produced knowledge are

equally important steps in the Knowledge Discovery Process. The mission of Bioinformatics as a new and critical research domain is to provide the tools and use them to extract accurate and reliable information in order to gain new biological insights. Data Mining in Structural Bioinformatics is used for the analysis and prediction of 3D structure of biological macromolecules using techniques based on information theory, Bayes theory, nearest neighbours, neural networks and integrated multiple sequence alignments.

Bioinformatics Text Mining is used to extract the information about genes, proteins and their functional relationships from text documents through Bioinformatics Text Mining (BTM).

The field of Biological Data Analysis and Mining is involves various areas which leads to organization, maintenance and the analysis of Biological data. In Biological Sequence Analysis even though there are various prediction algorithms but having some or the other flaws need is to modify and adapt techniques including neural networks, Bayesian classifiers, decision trees, and support vector machines, as well as multiple classifier systems. In gene Expression Analysis the problem of smaller amount of data but having large number of features can be solved by Clustering algorithms called subspace clustering and Bi-Clustering algorithm and even Hierarchical Clustering can be availed and further techniques such as bootstrap, repeated measurements [14], mixture model- based approaches, sub-sampling and others have been proposed to deal with the cluster reliability assessment Cluster validation quality can be based on predefined partitions of the data. In the paper we explored the use of data mining techniques which are very valuable in biological data analysis.

VI REFERENCES

[1]. Luscombe, N.M., Greenbaum, D. & Gerstein, M. , "What is Bioinformatics? A Proposed Definition and Overview of the Field". Methods of Information in Medicine, 40(4), 346-358, 2001.

[2]. Hunter, L., "Life and Its Molecules: A Brief Introduction". AI Magazine, 25(1), 9-22, 2004.

[3]. Tzanis, G., Berberidis, C., & Vlahavas, I., "MANTIS: A Data Mining Methodology for Effective Translation Initiation Site Prediction.", Proceedings of the 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, IEEE, Lyon, France, 2007

[4]. Houle, J.L., Cadigan, W., Henry, S., Pinnamaneni, A. & Lundahl, S., "Database Mining in the Human Genome Initiative. Whitepaper, Bio-databases.com", Amita Corporation. Available:http://www.biodatabases.com/ whitepaper.html, 2004.

[5]. Krallinger, M. & Valencia, A. "Text-mining and information-retrieval services for molecular biology". Genome Biology, 6(224),

[6]. Cohen, A.M. & Hersh, W.R., "A survey of current work in biomedical text mining". Briefings in Bioinformatics, 6(1), 57–71, 2005.

[7]. Ma, Q. & Wang, J.T.L. "Biological Data Mining Using Bayesian Neural Networks: A Case Study", International Journal on Artificial Intelligence Tools, Special Issue on Biocomputing, 8(4),433-451, 1999.

[8]. Zien, A., Ratsch, G., Mika, S., Scholkopf, B., Lengauer, T. & Muller, R.-K. "Engineering Support Vector Machine Kernels that Recognize Translation Initiation Sites". Bioinformatics, 16(9), 799-807, 2000

[9]. Zien, A., Ratsch, G., Mika, S., Scholkopf, B., Lengauer, T. & Muller, R.-K. "Engineering Support Vector Machine Kernels that Recognize Translation Initiation Sites". Bioinformatics, 16(9), 799-807, 2000.

[10]. Aas, K, "Microarray Data Mining: A Survey. NR Note", SAMBA, Norwegian Computing Center, 2001.

[11]. Smolkin, M. & Ghosh, D. "Cluster Stability Scores for Microarray Data in Cancer Studies", BMC Bioinformatics, 4, 36, 2003.

[12]. Tzanis, G. & Vlahavas, I. "Mining High Quality Clusters of SAGE Data. Proceedings of the 2nd VLDB Workshop on Data Mining in Bioinformatics", Vienna, Austria, 2007.

[13]. Xing, E.P., Jordan, M.I., & Karp, R.M.. "Feature selection for high-dimensional genomic microarray data", Proceedings of the 18th International Conference on Machine Learning, 601-608, 2001.

[14]. Yeung, Y.K., Medvedovic, M., & Bumgarner, R.E. "Clustering Gene-Expression Data with Repeated Measurements. Genome Biology", 4(5), R34, 2003.