A Study on Data Processing and Storage for Gene Expression

Alpna Sharma*

Abstract

Biomedical and Biological data is high intensity data that exists in different forms. This data is having the criticalities at each stage including the data generation, maintenance and processing. In this paper, an exploration of each aspect of gene expression data is defined. At first, the paper has explored each data form and elements of gene expression. The paper has also defined different organizational structured form of this larger data form. The preprocessing stage and transition of data to normalized form is defined. Finally, the paper has defined different methods of knowledge acquisition applied on gene data.

Keywords— Gene Expression, Structured DNA, Microarray, Clustering

INTRODUCTION

Biological data processing is the application area that inherits the molecular structure, genetic information, and phenotypic properties in composite and individual form. This data form is extracted from real environment and interpreted to generate various decisions and predictions. Each phase of information processing is associated with various challenges because of its representation, storage and the sequence form. Because of these challenges, this area provides wider scope for research.

---

*- Assistant Professor , Vivekananda School of Information Technology Vivekananda Institute of Professional Studies

The core element of this kind of information is proteins which are present in different organisms in different forms. The relational existence of these protein structures is defined by the multiple sequence composition. The organization, alignment and the search process on these structural patterns is required to extract the

abstracted information. This information exists in the form of sequence fragments as well as the gap between the fragments. The efficient information extension and the position assistive information can be used to describe the profile characteristics of living things in different forms. Based on the assumption that different spectral frequencies characterize different biological processes, more dynamic information representation is provided by Gene Expression (Chun-Pei, 2014; Bandyopadhyay, 2013; Mitra, 2006). It provides the cell function and behavior using multi dimensional features. In this section, different static and dynamic information forms and the sources are described.

DNA

The nucleus specific cell structure composition can be described as DNA molecule in a composite form. A DNA structure is described by the string of four nitrogenous bases called Adenine (A), Cytosine (C), Guanine (G) and Thymine (T). The genome is arranged into 24 different chromosomes which contains multiple genes. The functional and featured units of heredity are included in these genes. The region and structure analysis is required to analyze the relative features and the quantity of proteins. DNA sequence of human starts with TTCCTCCGCGA and comprise of about 10011 characters. This sequence is able to describe the characteristics and behavior of human beings_. To extract the information computational biology is required. These searches and processing operations can identify the required short sequences. These short sequences with locations are identified as regulatory regions. These repeated patterns are having higher significance to define some biological quality (Cheng-Pi, 2014; Fa, 2012; Mitra, 2006). DNA sequence processing is required to generate positional and repeated elements sub sequences.

Proteins

Protein is the basic organic component that generates amino acid to form the organic molecule. These molecules are formed using amine (NH) and carboxylic (CO) acid groups. These collectively form the side chain. Proteins are formed by using the linear chain of amino acid that actually forms a cell. The chemical features of these cells are different based on the unique structured chain formation. The cell functions and the descriptive properties are defined as the protein structure. The stable structure prediction can be done by observing the linear sequence of amino acids. The protein molecules can be formed in different ways by applying different sequence folding. The reaction and interaction of this amino acid and the molecule is required to be measured so that the secondary structure can be observed. These bonding and the structure patterns are observed to obtain the structural features. The structural conformation of these proteins is

determined by using the energy state observation. The protein folding can be minimized under bond length and the angle observation. The effect of this kind of bonding can be different based on the segment structure. The protein structures can be observed by predicting these relations and scores (Cheng-Pi, 2014; Bandyopadhyay, 2013; Cheng-Long, 2007; Gomaa, 2011).

Microarray

Microarray is the technology reform to define and store the gene expression at different level. These types of data are experimented under two main types called DNA and oligonucleotide arrays. These are defined with separate procedures and protocols. Microarray is defined as the solid surface which contains thousands of DNA molecules defined in the form of grid. Each of the grid cell itself describe the associated DNA sub-sequence. The array bounded and the measures are defined for these gene expression. Different samples are taken to hybridize the probes and generate more effective features (Chun-Pei, 2014; Nagi, 2011; Al-Timimi, 2004; Fa, 2012; Guo, 2007; Irsoy, 2012; Madeira, 2004). Gene expression levels are described from these samples which are taken as instances of biological processes or defined under some condition. These genes are defined on high dimensional vector of expression profile. The expression analysis is applied under the regulation mechanism. The sequence processing and the clustering are the major processes to filter and extract the hidden knowledge from these expressions. The biological role and the cluster validation based statistical perspectives are also defined to generate the segmented features. Microarray defines this collective technology to process these expressions at multiple levels and generate the abstracted knowledge.

Biological Networks

These networks can be defined under cell formation generating integrated complex reactions and its architectures. The node specific composition is formed where nodes are defined using genes. The interconnection or the links represent metabolic reaction of these genes when some processes are applied. The degree of each node is characterized by the relational aspects that generate a sequence or the pathway to describe the robust response with average pair analysis. The evolutionary topology and the environmental feature effect also scale up these properties and define a complex and logical systems. All the network operations like path generation, sub graph generation can be applied to generate the profile or the process information. The experimentation can be applied based on the gene regulatory network to provide more organized and coupled formation. The strength of the method is its high dimension support and easy

representation (Madeira, 2004; Mitra, 2006; Rastegar-Mojarad, 20IO). In this paper, a study work on different aspects of gene expression is presented. These aspects include forms of Gene expression, information acquisition methods from DNA Sequence and the knowledge extraction methods. In this section, the specification of biological information and data processing is described in details. The section has explored the different forms to represent the biological data. In section II, the work done on different forms by earlier researchers is described. In section III, the basic processing behavior applied on gene expression and microarray data is defined. The structural formulation is also described here. In section IV, different methods of knowledge generation and extraction are defined. The paper is concluded in section V.

REVIEW OF LITERATURE

Many researchers have published their work to explore the gene expression processing and associated data mining techniques on different datasets in different application areas. Some of the published works are discussed in this section. Knowledge acquisition from connected gene expression dataset, using mining techniques, is provided by Chun-Pi, 2014. Author applied Meta information processing with effective identification of biomarkers to generate the logic model. Author applied heuristic genetic method for process selection. The proposed method generated drug discovery for caner data based on effective knowledge processing and extraction. The core study on the gene regulatory network (Gomaa, 2011) is provided for health application domain to explore different diseases. Author summarized the constraints associated with the model and identified the challenges faced in topological processing. Author identified the intervention modeling to set the modeling under gene network specification. The comparative study (Guo, 2007) on different gene expression profiles was presented. The experimental behavior and the associated disease formulation for human, plants and animals are studied. The process description relative to the study is explored with suggestions for gene pattern generation and processing. Author setup the constraints to explore the criticality in various application domains. The similarities and challenges are also identified in different profiles. A study on various filtration methods for feature selection (Lazar, 2012) is provided to process analysis on gene expression. High dimension data was processed using different filters and gene prioritization method to discover the integrated patterns. A unified framework is defined to process different application domains with constraint specifications and with parameter adjustment.

Clustering (Al-Timimi, 2004) is also considered as the major tool for information acquisition. Different clustering algorithms can be applied individually and in combined form to generate data patterns. These algorithms can be analyzed under various measures and for different kind of validity processing to generate

effective results. Author (Fa, 2012) explored the significance of clustering for gene processing. He defines the feature specific reviews on five different clustering algorithms applied with validation specifications on benchmark datasets. Author identified the strengths and weaknesses of these algorithms. Daxin Jiang also provided the study work on cluster analysis. Author analyzed work on large datasets with specification of different functional genomics and interprets the challenges on large data forms. The complexity based analysis is identified under various measurements while working with different clustering methods. The structure patterns based feature exploration is provided to analyze the cluster. The aspect based validation problems are also identified to generate the promising trends from cluster processing. Biclustering (Madeira, 2004) algorithms are the activity processing based analytical processes to generate the data segments based on similarity observation. The conditions for cluster formation and dimension specifications are also defined to control the data processing with field name and information retrieval specification. Direct, associated and indirect clustering methods are applied in an integrated form to generate the optimized solution. Different Biclustering methods (Rastegar-Mojarad, 20IO) are explored by the author under non-supervised learning methods. The comparative observations related to these methods are provided by the author. Triclustering (Mahanta, 2011) is another clustering form to process with high dimension dataset to form 3D clusters. When the large feature descriptions are available then to expose the data issues in more clear form, the triclustering can be applied. The constraints specification and coherent relation also improves the clustering outcome. The probabilistic constraint(Romdhane, 2009) based unsupervised clustering is presented to process the vast data resources in disease identification. The behavior analysis based data patterns are generated to compute the cluster members. The generated more accurate clusters for effective gene expression processing.

DNA Microarrays(Al-Timimi, 2004) are the warehouse form of gene expression which can be applied and controlled by setting the prototypes conditions. The biological knowledge can be transited to gene expression based knowledge form using the biological information processing and provide the effective knowledge discovery. The design phase processing can be controlled by identifying the contribution of these knowledge aspects. (Bandyopadhyay, 2013) provided a comparative study on different processing and statistical tests on this data form. Author identified the parameters applied to control the test methodology and compared them for real time datasets. Author also expressed the various checks applied to acquire the knowledge and formulate them under data distribution and constraint generation. To perform the pattern classification(Chuang, 2007) on microarray data using neurofuzzy approach is defined. The feature study at primary level was process to present them in gene pattern form. Later on the fuzzy inference system is

applied with mining technique to predict the learned pattern. The intelligent optimization method has improved the accuracy of classification process.

The classification is the major task applied on bioinformatics patterns to take the predictive decisions. A study on classification method(Irsoy, 2012) and design modeling is provided. The experiment design and the constraint specific description of various statistical tests is provided by the author. The work also presented the analytical behavior of different performance measures along with exploration of guidelines and the associated methodology. Soft computing methods can be applied to generate the features in more exclusive form(Mitra, 2006). The pattern recognition and the classification can be regulated using these algorithms. The sequence processing, structure processing and association to the recognition and classification methods can be applied for effective generalization of data modeling. These methods reduce the efforts and improve the accuracy. Swierniak et. al. generated data patterns for class prediction using algebraic methods. Author improved the gene selection, pattern discovery and class prediction so that the more accurate decisions are carried out.

MICRO-ARRAY PROCESSING

To process on biological information, the first stage is to collect it from real means. The data must be collected with specification of nature of nucleotide bases. As the data is collected, it is raw form and it requires a series of processes called miniaturization and automation to apply the biological procedures over it. There are number of tools or the platform to organize this genomic data. These tools are having own characterization so that the high dimensional data will be stored with loss. The storage architecture and information extraction process are also defined by the platforms (Chuang, 2007; Fa, 2012; Irsoy, 2012; Madeira, 2004). As the data is collected and maintained, it is recognized as the particular gene profile. Each gene expression is a structured form with high dimensional data with thousands of specified genes. The time series specific large data is described with relative constraints and conditions. Some filtration process can be applied to index system so that the utilization effective data will be presented. This information can be defined in the form of a network or the high resolution images. The image data is the normalized data which can be processed and analyzed effectively. This normalized data is filtered, corrected and visual featured formed data. To normalize this data some standardization methods is applied with specification of reference format or structure (Nagi, 2011; Al-Timimi, 2004; Bandyopadhyay, 2013). The structured gene expression is shown in Figure 1.

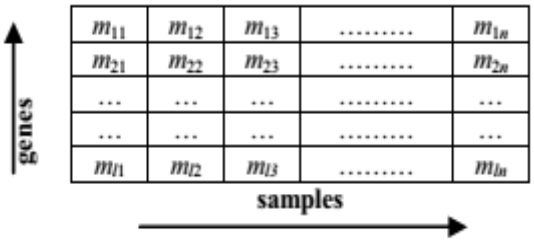

Figure 1: Structured Gene Expression

This dataset representation is in form of n x l matrix. Each of the element here represents the expression pattern with n genes G = g1,g2...gn. These gene set are presented by row and the column represents the profile with sample set S = s1,s2...sl specification. The conditions, constraints are also defined for each cell with specification f expression level. This kind of generated microarray is structured, easily formed and in normalized form. Error reduction and noise removal algorithms are also applied to generate the effective information.

Once the gene data is described, the pattern expression can be analyzed under similarity measure or partial similarity measures. The search methods are applied to identify the existence of particular pattern and its location. The functional similarity based expression pattern are called co-express genes. If two genes are regulated by similar transcription feature, then these expressions are applied and exist in same cellular process. Such kind of expressions is called co-relation. Another expression relation is known as coherent gene expression. This is characterized by multiple co-expressed genes. The template map with relative pattern is defined and the corresponding genes that follow the same template comes in this category. The divergence analysis and the regulations are also observed. The positional estimation, length and the repetitions are the main criteria based on which the pattern similarity or the partial similarity can be estimated. The process and characteristics driven mapping can be applied to generate the functional similarity.

Similarity Measures

The gene expression mapping from the group is done based on the pattern map to identify the existence or non-existence of some feature. Some similarity measures are required to measure this content and structured mapping. The high dimensional space can be analyzed under different properties map to generate the measures. Some of these similarity measures with relative constraints are listed here

1. Profile Distance analysis: This distance cannot be negative
2. For exact match, the distance between two genes must be zero, which is identified as perfect identical match.

3.The distance between two profiles can be measures under different distance methods including frequency measure, size measure, Euclidian distance measure, cosine distance measure etc.

The validity of these measures and constraints depends on the criteria of measuring including the point driven map, condition controlled map and the time scale based map. If the data is taken from real world, all the gene features must be map to identify the effective pattern occurrence.

KNOWLEDGE GENERATION METHODS

Once the gene data is stored in structured and normalized form without any integrated impurities, the next work is to apply different operations to generate the patterns and apply the knowledge acquisition. This knowledge formed transformation of gene expression can be done through some supervised or unsupervised learning methods. Different functional forms can be defined to process these expressions. Some of these processing forms are defined in this section.

Partitioning Method

It is the typical unsupervised learning method called clustering. In this method, the dataset is divided in smaller segments or partitions based on feature or the pattern type analysis. The method requires setting the centers of the partitions. The distance measures are applied on these centers and the data instances. The minimum distance or the maximum constraint map condition is applied to identify the cluster member. The method is defined for n number of iterations and in each iteration more clear partition is identified. Frequency, size or the sub-pattern ability can be considered as the parameter to perform distance estimation. The quantitative formed analysis is able to identify the relevant clusters. Each partition can describe the degree of relevancy of the mapping. The method also is able to identify the outlier or the data which is support to any available evidence or the constraint. Different similarity measure and the update data modes can be applied to generate the category of data. The conversions are also applied to formulate these clusters. These methods require the earlier knowledge of number of expected patterns that is not possible to identify while working with gene expression. The dimension of data is also a challenge to apply the clustering approach.

Hierarchical Method

To process high dimension data in layered form, a series of clustering methods can be applied in tree form. This nested clustering method is called hierarchical clustering approach. As the method defines, the re-filtration of any generated cluster can be further processed to acquire more information. This nested cluster

operation can be applied on each or the individual partition. The nested operation can also update the associated constraints and the conditions to gain more clear data partition. The data dependency and linkage is defined under criteria specification. The cluster distance and the complete linkage with maximum distance specification can be applied to generate the data existence in specific clusters. The scatter graph with employed links can be considered as the effective measures to decide the cluster members. The method is able to provide more sensitive and hidden information. It is able to manage high dimensional quantified information. The structural form of this method is more descriptive so that the easy representation of clusters can be obtained.

Density Based Method

These methods are applied on object space with dense area specification. The sparse clustering method is applied with conditional constraint specification to generate more effective clusters. The density based clustering method is applied to generate the arbitrary instance and specification of spatial data form is done. The method applies the neighborhood analysis with pattern effective map on specific regions. The partial clusters or the noise identification can be done using these methods. If some abnormal patters are required from spatial data structure, the data tracking and cluster ordering can be done. The probability set based grid map can be applied with distance factor specification. As the partial data is processed, the lesser data space is required which provide information that is more effective but as the size of data grows, the computation increases abnormally because of high level refactoring. Multiple process sequence with different constraints are applied to generate the abnormal information. The method is able to manage the high dimensional data and can recognize the outlier over the data space.

CONCLUSION

Biological Data Processing in the form of DNA Sequence and Gene Expression is identified as the core process model to generate the predictive decisions for Living Components. Earlier researchers defined work on clustering algorithms and feature generation method to provide effective knowledge acquisition and representation. In this paper, different aspects of biological information are defined. The paper identified different forms of this information including gene expression. The paper also explored the biological data storage and processing methods. Different knowledge generation methods applied on gene expression are also defined in this paper.

REFERENCES

Al-Timimi, A. (2004). Knowledge Discovery in a Microarray Data Warehouse, Proceedings of the International Conference on Information Technology: Coding and Computing (ITCC'04), Vol 2, pp 831-836

Bandyopadhyay, S. (2013). A Survey and Comparative Study of Statistical Tests for Identifying Differential Expression from Microarray Data, IEEE/ACM TRANSACTIONS ON COMPUTATIONAL BIOLOGY AND BIOINFORMATICS, 1545-5963@ IEEE. Vol. 11, Issue 1, pp. 95-115

Cheng-Pi, C. (2014). MiningABs: Mining Associated Biomarkers Across Multi-connected Gene Expression Datasets, BMC Bioinformatics, Vol 15, Issue 1, pp 173-183

Cheng-Long, C. (2007). A Neuro-Fuzzy Inference System to Infer Gene-Gene Interactions Based on Recognition of Microarray Gene Expression Patterns, *2007 IEEE Congress on Evolutionary Computation*, Singapore, 2007, pp. 904-910

Fa, R. (2012). CLUSTERING ANALYSIS FOR GENE EXPRESSION DATA: A METHODOLOGICAL REVIEW, Proceedings of the 5th International Symposium on Communications, Control and Signal Processing, ISCCSP 2012, pp 1-6

Gomaa, W. E. (2011). Modeling Gene Regulatory Networks: A Survey, *011 9th IEEE/ACS International Conference on Computer Systems and Applications (AICCSA)*, Sharm E. 2011, pp. 204-208

Guo, L. (2007). Comparison of gene expression profiles altered by comfrey and riddelline in rat liver, BMC Bioinformatics, Vol 8, Issue 7, pp 1-10

Irsoy, O. (2012). Design and Analysis of Classifier Learning Experiments in Bioinformatics: Survey and Case Studies, IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 9, no. 6, pp. 1663-1675*

Jiang, D. (2004) . Cluster Analysis for Gene Expression Data: A Survey, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, I*EEE Trans. on Knowl. and Data Eng.* 16, 11 (November 2004), pp 1370-1386

Lazar, C. (2012) . A Survey on Filter Techniques for Feature Selection in Gene Expression Microarray Analysis, IEEE/ACM Trans. Comput. Biol. Bioinformatics *9, 4 (July 2012), pp 1106-1119*

Mahanta, P. (2011) . Triclustering in Gene Expression Data Analysis: A Selected Survey*,* 2011 2nd National Conference on Emerging Trends and Applications in Computer Science*, Shillong, 2011, pp. 1-6*

Mitra, S. (2006) . Bioinformatics With Soft Computing, IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS—PART C: APPLICATIONS AND REVIEWS , *Vol* 36, Issue 5 (September 2006), pp 616-635

Nagi, S. (2011). Gene Expression Data Clustering Analysis: A Survey, *2011 2nd National Conference on Emerging Trends and Applications in Computer Science*, Shillong, 2011, pp. 1-12

Rastegar-Mojarad, M. (2010). A Survey on Biological Data Analysis by Biclustering, *2010 International Conference on Educational and Information Technology*, Chongqing, 2010, pp. V1-100-V1-103.

Romdhane, L. B.(2009). Mining microarray gene expression data with unsupervised possibilistic clustering and proximity graphs, *Applied Intelligence,* Vol 33, Issue 2 (October 2010), pp 220-231

Swierniak, A. (2009) . Class prediction and pattern discovery in microarray data – artificial intelligence and algebraic methods, *2009 First Asian Conference on Intelligent Information and Database Systems*, Dong Hoi, 2009, pp. 57-60