ISSN (Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization, Volume3, Special Issue 6, February 2014

National Conference on Emerging Technology and Applied Sciences-2014 (NCETAS 2014)

On 15th to 16th February, Organized by

Modern Institute of Engineering and Technology, Bandel, Hooghly 712123, West Bengal, India.

Graphical Model of Genes: A Review

Maity Sheuli, Maiti Krishna Kanta, Ghosh Sagarika

Assistant Professor, Dept. of CSE, Modern Institute of Engineering and Technology, Bandel, West Bengal, India

ABSTRACT : Gene is the regulator of the genotype and phenotype of an organism. Presence of one or more Gene mutation or some environmental factors result the variance in Expression level that creates Differentially Expressed Gene (DEGs), which can cause the predisposition or susceptibility to a particular disease. The most challenging issue for the modern geneticists is to discover and understand the function and variation of Genes interconnecting between themselves and other environmental factors; as well as to understand how such qualities affect health and disease. To differentiate between normal Gene and DEGs, researchers have suggested various Graphical Modeling approach using Probabilistic, Statistical and Graph theory methodology.

I. INTRODUCTION

Understanding biological systems with thousands of Genes would require organizing similar parts by their properties. Methods to group Genes with similar expression patterns have proved useful in identifying Genes that contribute to common functions or Genes that are likely to be co-regulated. The hypothesis that many human diseases may be accompanied by specific changes in Gene expression has generated much interest in Gene expression monitoring at the genome level.

To understand the function and variation of Genes interconnecting between themselves and other environmental factors; as well as to understand how such qualities affect health and disease, several powerful techniques have been developed, such as DNA Microarray [1], SAGE [2], BodyMap [3], MPSS [4] for collecting the Gene expression patterns. A major challenge in computational biology is to uncover Gene/protein interactions and biological pathways at the molecular level from such measurements, so that, the potential members of Gene groups responsible for specific physiological processes can be identified. In this context, Graphical models represent a combination of probability theory and graph theory that may provide a suitable tool to represent the dependence structures through a graph for multivariate random observations. Many statistical, probabilistic, functional models have been developed by measuring pair-wise measurements such as correlation, Euclidean distance, co-variance or mutual information between the Genes till now. In this paper, we, thereby, represent a survey on different Graphical Models of Genes and recent trend.

II.GENE

DNA (Deoxyribonuclic acid), the biomolecule that carries the genetic information, is two long-twisted strands made up of four similar chemicals called bases and abbreviated as A, T, C, and G that are repeated over and over in pairs. DNA can be copied into DNA i.e. ***DNA Replication***, DNA information can be copied into mRNA i.e. ***Transcription***, and proteins can be synthesized using the information in mRNA as a template i.e. ***Translation*** is termed as central dogma [5]. The most important phase is the protein synthesis because they are the enzymes that rearrange chemical bonds to carry signals to/from the outside of the cell and within the cell as well as regulate cell process, turn Genes on/off and control their rates.

A Gene is a sequence of nucleotides along a DNA strand - with '*start*' and '*stop*' codons and other regulatory elements which specifies a sequence of amino acids that are linked together to form a protein. Genes are passed down from parents to children conferring the traits to the offspring. Genes are organized and packaged in thread-like structures called the Chromosomes that make up the organism's whole genome and are stored in the nucleus of the cell, which is the master organelle for regulating cellular life processes and cell reproduction. Researchers have identified about 40,000 Genes in Human till now. The complete copy of the entire set of human Gene instructions is called as Genome.

ISSN (Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization, Volume3, Special Issue 6, February 2014

National Conference on Emerging Technology and Applied Sciences-2014 (NCETAS 2014)

On 15th to 16th February, Organized by

Modern Institute of Engineering and Technology, Bandel, Hooghly 712123, West Bengal, India.

Although the genome is the same in all somatic cells within an organism, cells are differentiated through differential Gene expression.

There are three postulates [6] of differential Gene expression. Firstly, the DNAs of all differentiated cells are identical. Secondly, the unused Genes in differentiated cells are not destroyed or mutated, and they retain the potential for being expressed. Lastly, only a small percentage of the genome is expressed in each cell, and a portion of the RNA synthesized in the cell is specific for that cell type. Common human diseases result from the interplay of many Genes and environmental factors. Therefore, a more integrative biology approach is needed to unravel the complexity and causes of such diseases by identifying the differences between healthy and affected tissues by forming specific regulatory networks that are dysfunctional in a given disease state. Although, research still have not reached a stage where the Elucidation of differential regulatory networks is commonly feasible. Recent advances have described the first steps towards this goal - the identification of differential co-expression networks from differential Gene expression and evaluate how this shift will affect the study of the genetic basis of disease.

III.GRAPHICAL MODEL

The vast quantity of genomic expression data generated by genomic expression arrays promises a significant opportunity on the understanding of basic cellular processes, the diagnosis and treatment of disease by transforming biology, medicine, and pharmacology using computational methods. Exploration of co-regulated Genes can identify potential members of Gene groups responsible for specific physiological processes. The various mechanism by which Cells control and regulate the transcription of their Genes, moreover Genome sequencing, Gene recognition, Gene/protein interactions and biological pathways at the molecular level are the recent trends in bioinformatics. Genes, may have same function, but not necessarily share similar transcriptional pattern [7]. Conversely, Genes having different functions can have a similar expression profile simply by chance or stochastic fluctuations. Also, Clustering cannot reveal functional relation among Genes with expression patterns that show very little correlations. To explain these patterns, there must have some postulate models describing the underlying biological mechanisms and then score these models in order to determine which are most consistent in reference the observed data. Recent research techniques produce a model-driven framework for the analysis of Gene expression data. It represents hypotheses about the characteristics of Genetic regulatory networks in a compact probabilistic form and can develop various effective methods for scoring these hypotheses in comparison with one another in terms of their relative ability to explain noisy expression data. Typically, analysis is performed by measuring pair-wise measurements such as correlation, Euclidean distance, co-variance or mutual information of Genes and results are visualized graphically. Model driven framework is used presently for the analysis purpose of Gene Expression data. The model may be drawn depending on segregation network (inheritance relationship), allele network (removing unnecessary segregation indicators and associated arcs), genotype network (Mendelian Inheritance) or phenotype network (Observable function and behavior). The Gene, allele, genotype or phenotype are represented as individual vertices and the relation between them are represented by directed edge (the edge $a \rightarrow b$ implies that Gene a regulates b in the sense that the expression of b is a direct consequence of the expression of a.). To form a graphical model, following steps [8] are maintained: **1.** Form an undirected graph by marrying the parents and moralizing the graph **2.** Triangulation by adding fill-in edges to the moral graph until all cycles involving more than three nodes have chords. **3.** Construct the junction tree by identifying the cliques. **4.** Loading the junction tree identifying the potentials in factorization, and **5.** Incorporation of observations.

IV. GRAPHICAL MODEL APPROACH

A. CONVENTIONAL MODEL

Conventional graphical modeling [9] depends on the Genetic regularity network where Genes are represented by the Vertices of a graph and conditional dependencies between their expression profiles are encoded by directed and undirected edges, with discretized and continuous data maintaining the **Global Markov Property** [7]. It becomes complicated when the number increases as it entails a large number of spurious edges in the model and also the evaluation of knowledge becomes difficult. The evaluation can be made easier by dividing the graph into smaller sub-

ISSN (Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization, Volume3, Special Issue 6, February 2014

National Conference on Emerging Technology and Applied Sciences-2014 (NCETAS 2014)

On 15th to 16th February, Organized by

Modern Institute of Engineering and Technology, Bandel, Hooghly 712123, West Bengal, India.

networks or can simply ignore the least important Genes. But for doing that, some important features may be missed. The prior knowledge about the Differentially-Expressed-Genes (DEGs) is required.

B. GRAPHICAL GAUSSIAN MODEL (GGM)

A solution to Conventional modeling with many Genes is that, condition will not be applied on all Genes at a time, instead, first modeling to small sub-networks with few Genes again and again and these sub networks are then combined for making inferences on the complete network. This modified graphical modeling approach also known as **co-variance selection model** makes it possible to include many Genes in the network while studying dependence patterns in a more complex and exhaustive way than with only pair wise correlation-based relationships.

- **Standard GGM**

The less important edges are reconsidered in Standard GGM. **Bootstrap Resampling method** is applied and a high cutoff level as 0.8 led to reasonably low number of selected edges. In such a high level of cut-off, many true edges may be missed. If conditions are applied simultaneously on many variables, it will introduce many spurious edges that have little absolute pair wise correlation but create a high absolute partial correlation into the model [10].

- **Modified GGM**

To improve upon the drawbacks of Standard GGM, Modified GGM has been introduced where the graph has been split into two sub graphs, each displaying the sub network of one module and its neighbors. Within a pathway many consecutive or closely positioned Genes are potentially jointly regulated. Two versions have been developed of this method: a ***Frequentist approach*** where each edge is tested for presence or absence; and a ***Likelihood approach*** with parameters θ_{ij}, which describe the probability for an edge between i and j in a latent random. But, it can only reveal linear dependencies between Genes [10].

In the multivariate t-distributions case, for more robust inference of graph Modified GGM is represented in a new look where the penalized likelihood inference combined with an application of the EM algorithm provides a computationally efficient approach to model selection. Two versions of multivariate t-distributions have been considered, one of which requires the use of ***approximation techniques***, using a Markov chain Monte Carlo EM algorithm based on a Gibbs sampler named ***t-lasso*** and ***a simple variational approximation*** named $\boldsymbol{t^{*}_{VAR}lasso}$ that makes the resulting method feasible in large problems [11].

C. *BOOLEAN MODEL* [12]

Interactions between mRNAs and proteins are described as logical (Boolean) functions such that the state of Genes is described by binary (ON/OFF) variables. The dynamic behavior of each variable, that is, whether it will be ON or OFF at next moment, is governed by a Boolean function. In this model, each mRNA or protein is represented by a node and the interactions between them are encoded as directed edges. A Boolean or logical function is written as a statement acting on the inputs using the logical operators "*and*", "*or*" and "*not*" and its output is $1(0)$ if the statement is true (false). It is assumed that Genes are equivalent, and their interactions form a directed graph in which each Gene receives inputs from a fixed number K of randomly selected neighbors. In a ***Random Boolean Network (RBN)*** [12] the functions governing the state of each node are randomly selected from the 2^{2^K} possible K-input Boolean functions, and kept fixed afterward. For $K > 2$ there are around N/e possible cycles whose length scales exponentially with N, however, for $K = 2$ both the number and length of the limit cycles is only$\sqrt{N}$.

D. BAYESIAN MODEL

A Bayesian network is a graph-based model of joint multivariate probability distributions that captures properties of conditional independence between variables in case of noisy observations [13]. This representation consists of two components. The first component, E, is a directed acyclic graph (DAG) whose vertices correspond to the random variables $(X_1, X_2, \dots, X_n)$. The second component, θdescribes a conditional distribution for each variable, given its

ISSN (Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization, Volume3, Special Issue 6, February 2014

National Conference on Emerging Technology and Applied Sciences-2014 (NCETAS 2014)

On 15th to 16th February, Organized by

Modern Institute of Engineering and Technology, Bandel, Hooghly 712123, West Bengal, India.

parents in G. Together, these two components specify a unique distribution on $(X_1, X_2, \dots, X_n)$. In this model, the graph G encodes the Markov Assumption: Each variable X_i is independent of its non-descendants, given its parents in G. Satisfying this property, the conditional joint distribution for each variable X_i can be decomposed in the following form:

$$P(X_1, X_2, \dots, X_n) = \prod_{i=1}^{n} \quad P(X_i | Pa^G X_i)$$

Where, $Pa^G X_i$ denotes the set of parents of X_i , that can be Discrete or Continuous.

Theorems: Two DAGs are equivalent if and only if they have the same underlying undirected graph and the same v-structures (i.e. converging directed edges into the same node, such as $a \longrightarrow b \longleftarrow c$) [14].

E. S-SYSTEM MODEL

Among the dynamic and continuous approaches is the '***S-system***' [15] [16] is a type of power-law formalism and is based on a particular type of ordinary differential equation in which the component processes are characterized by power-law functions. As it allows the customizing of analytical and computational methods, it has a great advantage in terms of system analysis and control design. Moreover, using S-system parameters, steady-state evaluation, control analysis, and sensitivity analysis of a given system can be established mathematically [17]. On the other hand, the S-system has a major disadvantage that all of its large number of parameters i.e. 2n(n + 1), where n is the number of state variables like concentration must be estimated which often causes bottlenecks problem. ***Real-coded Genetic Algorithm*** [18] has been used over Conventional binary GA to determine many parameters simultaneously with high accuracy and high optimization speed. For numerical integration of the S-system, there is a high-speed algorithm called **'*Evaluation and Simulation of Synergistic Systems (ESSYNS)*'** [19].

F. SVM MODEL

In post-genomic biology, it is an important problem to predict the targets of a transcription factor (TF) by identifying subtle relationships between their expression profiles. The ***Support Vector Machine (SVM)*** [20] is a standard supervised machine-learning algorithm, based on recent developments in statistical learning theory [27] and models based on SVM is represented to determine the relationship between TFs and their targets. In this model, the pair is made between TF (say, R) and the Target (say T) and the directed edge $R \longrightarrow T$ denotes that transcription factor R regulates Gene T. In this approach, some ***Positive*** (i.e. only sequence-specific TFs) and some ***Negative*** (i.e. no regulatory relationship) examples needed to train properly. When there is a large difference between positive and negative examples of the data there occurs the ***Imbalance Problem*** [21] [22] which can be avoided by increasing the size of the under-represented set by random resampling and decreasing the size of the over-represented set by random removal of its members.

G. HIDDEN MARKOV MODEL

The statistical model, also called as **'Genie'**, that provides the framework for describing the grammar of a legal parse of a DNA sequence [24]. It provides simple solutions for integrating cardinality constraints reading frame constraints "***indels***" and homology searching, recently used in pattern recognition [25] and identification of Gene structure in *E. coli [26]*. In this model, edge represents the states in the state machine and nodes represent the transition between states. Let M=Model and a candidate DNA sequence, $X = \{X[1], X[2], \dots, X[n]\}$ is given, then the predicted Gene structure is defined as the ordered set of states, $\phi = \{q_1, q_2, \dots, q_n\}$, called the parse such that the probability of generating X according to ϕ is maximal over all possible parses. The current implementation of Genie [23] as follows: 1. **Length Distributions** (generates length histogram from the training sets), 2. **Splice Site Model** (a backpropagation feed forward network is trained with one layer of hidden units), 3. **Intron Model** (windowed null model), and 4. **Exon Model** (the GC-content, codon usage and previous codon are simply integrated in a single discriminator).

V.CONCLUSIONS

In this review paper, we have studied and reviewed the different modeling approach. The different methods have estimated and also postulated different methodologies towards the implementation of graphical model of Gene. In

ISSN (Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

An ISO 3297: 2007 Certified Organization, Volume3, Special Issue 6, February 2014

National Conference on Emerging Technology and Applied Sciences-2014 (NCETAS 2014)

On 15th to 16th February, Organized by

Modern Institute of Engineering and Technology, Bandel, Hooghly 712123, West Bengal, India.

Conventional Modeling approach, the author mainly given trace to the Global Markov Property whereas Gaussian Graphical Modeling approach represents more robust through co-variance selection of Genes. In this approach, the network model is subdivided into small subnetworks and each subnetwork is then processed through methods like *Frequentist* and *Likelihood* approach instead of evaluating the whole network as Conventional modeling. The Boolean modeling approach, described by R´eka Albert [12], plays an important role in the stability of the segment polarity Genes. Bayes theorem has been developed on Gene expression data to develop a Bayesian Model of Gene. The Graphical Model is also implemented through Genetic Algorithm by Jonikow and Michalewicz [18] to remove the bottleneck problem in case of S-System Modeling. Moreover statistical methods such as SVM, Hidden Markov Model is also used for this purpose. It is not possible to identify a particular most effective model for overall analysis of all human DEGs or normal Genes. Though Conventional method plays well at pathway-finding, the future research focuses more on dynamic changes and interactions among biological objects.

ACKNOWLEDGEMENTS

The authors acknowledge the support provided by Modern Institute of Engineering and Technology, Bandel.

REFERENCES

[1] Duggan, D.J., Bittner, M., Chen, Y., Meltzer, P., and Trent, J.M., *Expression profiling using cDNA microarrays*, Nature Genet., 21:1014, 1999.
[2] Velculescu, V.E., Zhang, L., Vogelstein, B., and Kinzler, K.W., *Serial analysis of gene expression*, Science, 270: 484487,1995.
[3] Okubo, K., Hori, N., Matoba, R., Niiyama, T., Fukushima, A., Kojima, Y., and Matsubara,K., *Large scale cDNA sequencing for analysis of quantitative and qualitative aspects of gene expression*, Nature Genet., 2:173179, 1992.
[4] Brenner, S. et al., *Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays*, Nature Biotech., 18:630634, 2000.
[5] Pierre Baldi and Sren Brunak, *Bioinformatics: The Machine Learning Approach*.
[6] *Developmental Biology*, Gilbert SF. Sunderland (MA): Sinauer Associates; 6th edition, 2000.
[7] Anthony Almudevar, Lev B. Klebanov, Xing Qiu, Peter Salzman, and Andrei Y. Yakovlev, *Utility of Correlation Measures in Analysis of Gene Expression*. The American Society for Experimental NeuroTherapeutics, Inc., Vol. 3, 384 395, July 2006.
[8] Steffen L. Lauritzen, Nuala A. Sheehan, *Graphical Models for Genetic Analyses*, 2003.
[9] Steffen L. Lauritzen and Nuala A. Sheehan, *Graphical Models for Genetic Analyses*, September 23, 2003.
[10] Anja Wille, Philip Zimmermann et al., *Sparse graphical Gaussian modeling of the isoprenoid gene network in Arabidopsis thaliana*, Genome Biology2004, 5:R92, 2004.
[11] Michael Finegold and Mathias Drton, *Robust Graphical Modeling of Gene Networks Using Classical and Alternative T -Distributions*, Institute of Mathematical Statistics, 2011.
[12] Reka Albert, *Boolean Modeling of Genetic Regulatory Networks*, Department of Physics, Pennsylvania State University, University Park, PA 16802, USA.
[13] Nir Friedman, Michal Linial, Iftach Nachman, Dana Pe'er, *Using Bayesian Networks to Analyze Expression Data*, Fourth Annual International conference on Compu-tational Molecular Biology, 2000.
[14] J. Pearl and T. Verma, *A Theory of Inferred Causation, In Principles of Knowledge Representation and Reasoning: Proceedings of the Second International Conference*, ed. By J. Allen, R. Fikes, and E. Sandewall, Morgan Kaufmann, San Mateo CA, 1991.
[15] Shinichi Kikuchi, Daisuke Tominaga, Masanori Arita,Katsutoshi Takahashi and Masaru Tomita, *Dynamic modeling of genetic networks using genetic algorithm and S-system*, September 28, 2002.
[16] Savageau, *Biochemical System Analysis: a Study of Function and Design in Molecular Biology*. Addison-Wesley, Reading, MA., 1976.
[17] Voit,E.O. *Computational Analysis of Biochemical Systems*. Cambridge University Press, 2000.
[18] Jonikow,C.Z. and Michalewicz,Z. *An experimental comparison of binary and floating point representations in genetic algorithms. InProceedings of the International Conference on Genetic Algorithms*. pp. 3138, 1991.
[19] Irvine, and Savageau, *Efficient solution of nonlinear ordinary differential equations expressed in S-system canonical form*. SIAM J. Numer. Anal.,27, 704735, 1990.
[20] Jiang Qian, Jimmy Lin, Nicholas M. Luscombe, Haiyuan Yu and Mark Gerstein, *Prediction of regulatory networks: genome-wide identification of transcription factor targets from gene expression data*.
[21] Japkowicz,N. *The class imbalance problem: significance and strategies*. Proceedings of the 2000 International Con-ference on Artificial Intelligence,1, 2000.
[22] Japkowicz,N. and Stephen,S. *The class imbalance problem: a systematic study*. Intelligent Data Analysis,6, 429450, 2002.
[23] David Kulp, David Haussler, Martin G. Reese, Frank H Eeckman, *A Generalized Hidden Markov Model for the Recognition of Human Genes in DNA*.
[24] Stormo, G. D., and Haussler, *Optimally parsing a sequence into different classes based on multiple types of information*. In ISMB-94 Menlo Park CA: AAAI/MIT Press, 1994.
[25] Rabiner, L. R. and Juang, B. H. *An introduction to hidden Markov models*. IEEE ASSP Magazine 3(1):4-16, 1986.
[26] Krogh, A.; Mian, I. S.; and Haussler, D. *A Hidden Markov Model that finds genes in E.coli* DNA. NAR 22:4768-4778, 1994.
[27] Vapnik,V. (1998)*Statistical Learning Theory*. Wiley, New York.