VISUAL DATAMINING OF BIOLOGICAL NETWORKS::TEMPORAL MODELLING OF A GENE NETWORK

Harishbabu.Kalidasu[1], Dr. Gudipati Murali[2]

[1]Research Scholar, [2]Research Supervisor, Acharya Nagarjuna University, Guntur, Andhra Pradesh, (India)

ABSTRACT

High-throughput tentative protocols have been exposed thousands of relationships in the middle of genes and proteins under different conditions. These reputed associations are being insistently mined to work out the structural and functional architecture of the cell. One functional tool for exploring this information which has been computational network analysis. In this paper, we plan a collection of new algorithms to survey the structure and evolution of large, noisy, and thinly (lightly) annotate biological networks.

To make easy of the study of earliest networks, we start a framework (call "network archaeology") for reconstructing the node-by-node and edge-by-edge arrival record of a network. While starting with a present-day network, we are applying a probabilistic growth of model backwards in time to find a high-likelihood before states of the graph. This allows us to discover how interactions and modules may have been evolved over time. While experiments in real-world social and biological networks, we find that particular algorithms can be improve considerable features of ancestral networks that have long since left.

Our work is motivated by the need to understand large and compound biological systems that are being exposed to us by unsatisfactory data. As the data continues to move in, we believe that computational network breakdown (analysis) will continue to be a necessary tool towards this end.

I. INTRODUCTION

1.1 Human Perception and the Digital World

Nowadays, everybody speaks about digitalization. From the technical top of the view, it is very efficient, since electronic circuits work well on digital information. But the digital information is not for individual observation because regularly it does not point out a valuation. We do not know what it actually means if we vary a digit from 0 to 1. It is only a small change regarding its visual form, but the change can be enormous depending on the position of the digit.

Recognition of digital information takes more time. The history of digital watches is a good example. They were only fashionable for a short time due to the human perception method. It is easy to verify that analogue signs are better recognized because they indicate an evaluation. Digital communication is not only disappearing for watches, the car and aeroplane cockpits also provide good examples.

We advise a new generation of visual information system, called DataScope. This has a number of properties: digital information is translated into analogue one; queries, an necessary function of databases, can be realised; several features can be examined at the same time and only the visualisation capacity of the computers or human perception can limit it; comparison (relation) can be accomplished, i.e. the relation of two or more alternatives can be visualised at the same time.

[1]For the existence of detailed curation of technical (methodological) literature and increasingly consistent computational predictions have resulted in a formation of huge databases of protein interaction data. Over the years, these repositories have become a basic framework in which experiments are analyzed and latest directions of research are explored. Now the most broadly used protein-protein interaction databases and methods they employ to assemble, combine and expect interactions. We furthermore spot out the trade-off between comprehensiveness and accuracy and the main drawback; the scientists have to be aware by adopting the protein interaction databases in any single-gene or genome-wide analysis.

The repetitive cost decrease of high-throughput experiments and the growth of computational prediction methods have produced huge number of protein-protein interactions (PPIs). This facility to provide moderately broad and consistent sets of PPIs prompted the development of many databases aiming to gather round and combine the available data, all with a dissimilar focus and different strengths.

The PPI resources are now usually used for data analysis, data interpretation, and hypothesis testing. A complete list of more than 300 pathways and interaction databases is available from pathguide. However, the scope to which any of the available PPI datasets reflect the biological interactome is indefinite, and it is thus basically to vigilantly calculate the advantages and drawbacks of every interaction data source before using them. Definitely till now, no one can capture the full complexity of biological systems with a different protein variants, modifications, and spatial and temporal dependencies.

The importance of network visualization has been commonly known and most databases offer some type of network view; if not as a native use work in a browser, and then only if the network as a downloadable files that can easily be imported into visualization tools such as cytoscape[2] or NAViGaTOR[3]. These tools and some databases also permit users to layout networks, annotate nodes, and perform various types of network analysis such as clustering and term enrichment analysis.

II. BACKGROUND INFORMATION AND SOME STUDY ON DATAMINING

Among the quick (rapid) growth of computer and information technology in the most recent several decades, a huge amount of data in science and engineering has been always be generated in enormous level, moreover being stored in huge storage devices or elegant into and out of the system in the outline of data streams.

It has been widely known that the rapid development of computer and information technology in the last twenty years has primarily changed almost every field in science and engineering, transforming various disciplines from data-poor to ever more data-rich, and mission for the development of new, data-intensive methods to perform research in science and engineering.

Data mining, as the union of various intertwined disciplines, including *statistics*, *machine learning*, *pattern recognition*, *database systems*, *information retrieval*, *World-Wide Web*, *visualization*, and *many application domains*, has through enormous growth in the past decade.

To make sure that the advances of data mining research and technology will efficiently promote the evolution of science and engineering, it is essential to observe the challenges on data mining posed in data-intensive science and engineering and investigate how to added the features for developing the technology to make easy of new discoveries and advances in science and engineering.

III. INFORMATION NETWORK ANALYSIS

The Scientists (Users / People) habitually take care of a database as a data repository that stores huge sets of data and supports indexing, retrieval, and different kinds of updating and difficult query processing. On the other hand, entities/objects in databases are not inaccessible tuples; they have rich, inter-related semantic information that can be and there should be scientifically explored. One significant information that the most earlier research has not paid much concentration is to facilitate objects in databases are inter-related and linked with the help of foreign keys, across various relations or entity sets, forming huge information networks. Information network analysis methods can be steadily developed for in-depth network-oriented data mining and analysis, which is distant away from the scope of usual search functions provided in database systems.

With the development of Google and other useful web search engines, information network analysis has become an important research frontier, with broad applications, such as social network analysis, web community discovery, terrorist network mining, computer network analysis, and network intrusion detection. However, information network research be supposed to go away from explicitly formed, homogeneous networks (*i.e.*, web page links, computer networks, and terrorist e-connection networks) and explore intensely into *implicitly formed*, *heterogeneous*, and *multidimensional* information networks. Science and engineering present us loaded (various) opportunities on searching of networks in this path.

At present we have a group of huge natural, technical, social, and information networks in science and engineering applications, such as gene, protein, and microarray networks in biology, highway transportation networks in civil engineering, topic or theme-author-publication-citation networks in library science, and wireless telecommunication networks among commanders, soldiers and supply lines in a battle field. Within their information networks, each node or link in a network contains *valuable, multidimensional information,* such as textual contents, geographic information, traffic flow, and other properties. Furthermore, these type of networks could be highly *dynamic, evolving, inter-dependent* and *mutually supporting*.

IV. KNOWLEDGE DISCOVERY, UNDERSTANDING, USAGE OF PATTERNS

Scientific and engineering applications often handle massive data of high dimensionality. The goal of pattern mining is to find item sets, subsequences, or substructures that become visible in a data set through frequency not less than a user-specified threshold. Pattern analysis can be a valuable tool for finding correlations, clusters, classification models, sequential and structural patterns, and outliers.

"Knowledge Discovery is the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data"

Among the importance on collecting data ever-increasing around the world, there is a necessary need for a new generation of different techniques, methods and algorithms give support to researchers, analysts, decision makers and managers for extracting useful patterns from the fast growing volumes of data. These techniques and tools are the subject of the emerging field of knowledge discovery in databases (KDD). KDD has evolved from interaction and cooperation among such different fields as machine learning, pattern recognition, database, statistics, artificial Intelligence, knowledge representation, and knowledge acquisition for intelligent systems. The main idea in KDD is to discover a high level knowledge (abstract knowledge) from lower levels of relatively raw data, or to discover a higher level of interpretation and abstraction than those previously known.

Earlier 5 decades the concept of finding or discovering useful interesting patterns in data which has been addressed by different research groups and researchers, So, here how KDD relates to these other approaches. Such approaches have been given different names, such as exploratory data analysis, information discovery, information harvesting, data archaeology, and data pattern recognition. KDD applies machine learning and pattern recognition techniques to extract patterns implicit in a database. The new way of KDD addresses the overall process of discovering useful knowledge from data while data mining, statistic analysis and other such techniques address only a particular step in this process. KDD seeks incrementally to understand, to adapt and apply these patterns to upcoming cases or data sets. KDD uses statistical methods, particularly exploratory data analysis methods, but it sees their use as only one part of a more complete knowledge discovery process. Statistical methods and algorithms propose (suggest / offer) accurate methods for quantifying natural inferential doubts. KDD systems embed particular statistical procedure for and modeling data, evaluating hypotheses and handling noise inside an overall knowledge discovery framework. KDD approaches and methods are engaged on model extraction or construction or discovery, relatively than the factors to estimation of earlier hypothesize models. They manage best in the background of huge (large) sets with rich data structures. For such huge data sets, interpretations may already exist, coming from a particular field of analysis, by shifting the window of concern to a different aspect of that data base, we might obtain some new pattern for another purpose.

While comparing with traditional data analysis, the KDD process is interactive and iterative. One has to make a number of decisions in the process of KDD. In the following we explain the major steps in KDD.

- Understanding the domain knowledge and identifying the goal of KDD.
- Creating the main data set for the purpose of the KDD.
- Data Cleaning and Data processing: we need basic operations to eliminate noise from the data and to check the validity of the data. Data cleaning helps in the level of confidence in data analysis.
- Finding interesting features in the database.
- Apply a data-mining algorithm such as clustering, classification, regression, in ways which match with the original goal of the KDD process.
- Data mining and searching for interested patterns in the data.

- Interpretation of the result or found pattern. A good clustering or classifying approach should explain the result of such an approach. If a result cannot be interpreted properly, it may not be useful for further purposes.
- Prepare a new set of knowledge for future analysis, utilization or discovery purposes.

V. STREAM DATAMINING

Data Stream Mining[4] is the process of extracting knowledge structures from continuous, rapid data records. A data stream is an ordered sequence of instances that in many applications of data stream mining can be read only once or a small number of times using limited computing and storage capabilities. Examples of data streams include computer network traffic, phone conversations, ATM transactions, web searches, and sensor data. Data stream mining can be considered a subfield of data mining, machine learning, and knowledge discovery.

In many data stream mining applications, the goal is to predict the class or value of new instances in the data stream given some knowledge about the class membership or values of previous instances in the data stream. Machine learning techniques can be used to learn this prediction task from labeled examples in an automated fashion. Often, concepts from the field of incremental learning, a generalization of Incremental heuristic search are applied to cope with structural changes, on-line learning and real-time demands. In many applications, especially operating within non-stationary environments, the distribution underlying the instances or the rules underlying their labeling may change over time, i.e. the goal of the prediction, the class to be predicted or the target value to be predicted, may change over time. This problem is referred to as concept drift[5].

Stream data refers to the data that flows into and out of the system like streams. Stream data is usually in vast volume, changing dynamically, possibly infinite, and containing multi-dimensional features. Typical examples of such data include audio and video recording of scientific and engineering processes, computer network information flow, web click streams, and satellite data flow. Such data cannot be handled by traditional database systems, and moreover, most systems may only be able to read a data stream once in sequential order. This poses great challenges on effective mining of stream data.

Stream data is often encountered in science and engineering applications. It is important to explore stream data mining in such applications and develop application-specific methods, *e.g.*, real-time anomaly detection in computer network analysis, in electric power grid supervision, in weather modeling, in engineering and security surveillance, and other stream data applications.

VI. VISUAL DATA MINING

A picture is worth a thousand words. It has been numerous data visualization tools for visualizing various kinds of data sets in massive amount of data and of multidimensional space. In addition to that popular bar charts, pie charts, curves, histograms, quantile plots, quantitle-quantile plots, boxplots, scatter plots, there are also many visualization tools using geometric (*e.g.*, dimension stacking, parallel coordinates), hierarchical (*e.g.*, treemap), and icon-based (*e.g.*, Chernoff faces and stick figures) techniques. Moreover, there are methods for visualizing sequences, time- series data, phylogenetic trees, graphs, networks, web, as well as various kinds of patterns and knowledge (*e.g.*, decision-trees, association rules, clusters and outliers). There are also visual data mining tools that may facilitate interactive mining based on user's judgement of intermediate data mining results.

Figure Shows Sample Text Document Related to Information Visualization

Visual data mining as an art and science of testing meaningful insights out of large quantities of data that are incomprehensible in another way requires consistent visual data representations [information visualization models]. The frequently used expression “the art of information visualization” appropriately describes the situation. Though substantial work has been done in the area of information visualization. It is still a challenging activity to find out the methods, techniques and corresponding tools that support visual data mining of a particular type of information. The comparison of visualization techniques across different designs is not a trivial problem either. This research presents an attempt for a consistent approach to formal development, evaluation and comparison of visualization methods. Visual data mining is the advanced way of representing information over enterprise level application they also help in dealing with flood of information. Now a day’s visual representation made easy for analysis and classification of data. Many analysis algorithms like BIDE for frequent item and closed item sets, and classification techniques like SVM are proposed which represents static representation of data.

The main advantage of visual data exploration is that the user is directly involved in the data mining process, through analysis the results of the information visualization, user can integrate the specialist knowledge with various data mining algorithms. This research involves in about visual data mining techniques are analyzed combining with some national advanced data mining tools. The model of semantically organized place for data exploration can be useful for the development of computer support for visual information querying and retrieval in collaborative information filtering. The development of a representational and computational model of selected metaphors for data visualization will assist the design of virtual environments, dedicated to visual data explorations. The formal approach presented is based on the concept of semantic visualization defined as a visualization method, which establishes and preserves the semantic link between form and function in the context of the visualization metaphor. Establishing a connection between form and functionality is not a trivial part of the representing data graphically, whether the data consists of numbers of text, is not a straightforward procedure as numbers and text descriptions do not have a natural visual representation. This research involves in representing data visually has a powerful effect on how the structure and hidden semantics in the data is perceived and understood. Visual data mining can help in dealing with the flood of information. The advantage of visual data exploration is that the user is directly involved in the data mining process, through analysis the results of the information visualization, user can integrate the specialist knowledge with the data mining algorithm. We summarizes current visualization methods applied in data mining. Current applications about visual data mining technique are analyzed combining with some national advanced data mining tools. Trends are clarified based on the task and object of visual data mining.

VII. BIOLOGICAL DATA MINING

The fast progress of biomedical and bioinformatics research has led to the accumulation and publication (on the web) of vast amount of biological and bioinformatics data. However, the analysis of such data poses much greater challenges than traditional data analysis methods[6]. For example, genes and proteins are gigantic in size (*e.g.*, a DNA sequence could be in billions of base pairs), very sophisticated in function, and the patterns of their interactions are largely unknown. Thus it is a fertile field to develop sophisticated data mining methods for in-depth bioinformatics research. We believe substantial research is badly needed to produce powerful mining tools in many biological and bioinformatics sub fields, including comparative genomics, evolution and phylogeny, biological data cleaning and integration, biological sequence analysis, biological network analysis, biological image analysis, biological literature analysis (*e.g.*, PubMed), and systems biology. From this point view, data mining is still very young with respect to biology and bioinformatics applications. Substantial research should be conducted to cover the vast spectrum of data analysis tasks.

VIII. DATA MINING FOR SOFTWARE ENGINEERING

Software program executions potentially (*e.g.*, when program execution traces are turned on) produce huge amounts of data. However, such data sets are slightly different from the datasets generated from the nature or collected from video cameras since they represent the executions of program logics which is coded by the individual programmers. It is significant to mine such data to observe program execution status, improve system performance, isolate software bugs, detect software plagiarism, analyze programming system faults, and recognize system malfunctions.

Data mining for software engineering can be partitioned into static analysis and dynamic/stream analysis, based on whether the system can gather traces in advance for post-analysis or it must respond at real time to hold the online data.

Different methods have been developed in this domain by integration and extension of the methods developed in machine learning, data mining, pattern recognition, and statistics. For example, statistical analysis such as hypothesis testing approach can be performed on program execution traces to isolate the locations of bugs which decide program success runs from failing runs. Despite of its limited success, it is still a rich domain for data miners to research and further develop sophisticated, scalable, and real-time data mining methods.

IX. PROTEIN-PROTEIN INTERACTION

Proteins are the one of the most important molecule groups for living organisms / cells. These cells serve as enzymes for catalysis of metabolic process, hormones substances (signaling), structural or mechanical material, or from other substances (oxygen), The primary structure of a protein is a long sequence out of essentially twenty different amino acids connected by peptide bonds.

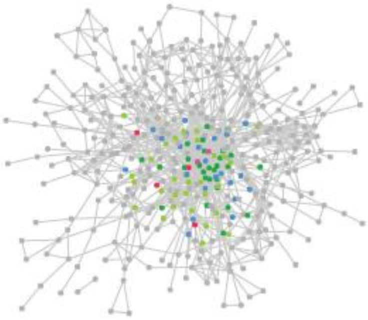


Fig a: Protein-Protein Interaction Network.

While a protein can interact with another protein (see fig: a), it means that to build a protein complex. The Protein protein interactions form large networks. The visualization of biologists in investigative the role of proteins and also gaining new insights about the processes within and across cellular processes and compartments, i.e., for formulating and experimentally test the specific hypothesis about the particular gene function.

While the existence of an interaction between two proteins is known, however the interaction type, where as activation, binding to, leftovers unknown. For the understanding of the biological processes, the data about the interaction type is very important, even if databases contain small information about those genes. Here we define a protein-protein interaction network as a directed graph G= (V,E,τ) where V is a set of proteins, E the set of directed interactions and τ: E-> T which defines the type of a each edge.

The representation of the networks has some relationships. While in a biological background, number of different types of relationships can be deliberated; those are physical interactions between proteins or genetic interaction exposed by combination of mutations. Once a large collections of diverse relationships are generated from a number of different high-throughput experimental analyses of a single biological system, while network visualization and their analysis can prove particularly very helpful[7].

X. NETWORK VISUALIZATION

Network Visualization contains two main aspects: the visual representation of nodes and edges (fig: b), and then design the basic graph. The visual representation of nodes and edges is very useful for producing assured information about the pathway. The design of the graph is important to expose the basic topology of the network. We concentrate on the design of the network.

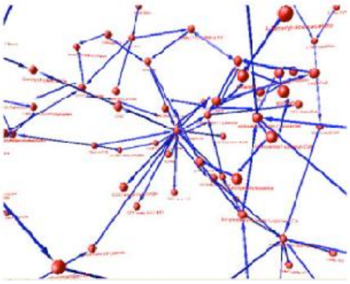


Fig: b: Network Visualization of Nodes and Edges

In few years ago, a large number of graph drawing algorithms and some methods are developed to automatically layout [8]node-edge diagrams. Some of the algorithms which include tree drawing algorithms force directed methods, multi-dimensional scaling methods and spectral graph drawing methods. For these details check the recent proceedings of graph or drawing and information visualization conferences.

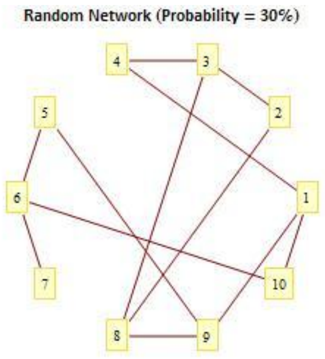
The network analysis by visual representation is rapidly becoming a popular technique t o clear understand about the large and complex networks. For a smaller networks, numerous complicated and feature applications have been developed. However, a network size approaches a number of available screen pixels, and few new methods will be needed to efficiently review the graph complexities[9].

10.1 Connection Patterns from Networks

10.1.1 Random Network

The network architecture that is opposite to the "regular" architecture is the homogeneous random network that was analyzed by Paul Erdös and Alfred Rényi in the mid twentieth century.

A random network is a hypothetical type of construct contains links that are completely random with equal probability, It is considered to be very extremely disordered. While using a random number of generator one should assigns links from one node to another node. Naturally random links results shortcuts to remote nodes. In fact, restrict the path length to else distant nodes. For example, links between node 6 and 10 or node 4 and 1 serves to reach clusters on the otherside of the network. This restriction of path length tends to increase the connectivity between the nodes.

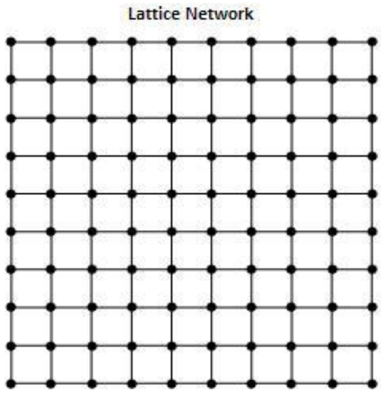


(Diagram Source: from Pattern signature)

According to the Stigmergic Systems "The advantage of a random network over a normal network is that, atleast there are few common links with nodes but the number of links between the nodes is very small. In fact many nodes there are in the network."

10.1.2 Regular Network

Regular networks are “Regular” because each and every node has exactly same number of links. These are highly ordered. In regular networks, a square “regular” framework like matrix is a non-random regular network where each node connect to all of its nearest nodes. This framework also be represented as rings, stars and trees. It can characterize schools, universities, groups, whereas the behavior of individual node which depends upon the behavior of the nearest nodes. While in the framework, the topological rule of each node is linked with all of nearest four nodes. No rule is needed to arrange or to define the framework degree distribution because the number of degrees (4) for each node is same.



(Diagram Source: from Patternsinnature)

While the random networks and regular networks are two general connection patterns in modeling of biological networks. This type of biological modeling network is not applicable to the network where their connection patterns exist between these two extremes (watts and strogatz 98). Generally some algorithms are developed to merge the properties of regular network and random network done over the changing pattern of randomly connected vertices in the network. As we see the erdox-renyi model (erdos and renyi 1996) will have some of the properties of random network and regular networks. These properties which consists of high clustering co-efficients of regular network and small world property of random network. The erdos-rengyi model explains it has small cultural coefficients and small characteristic path lengths, whereas the watts-strogatz model has very high clustering coefficient and small characteristic path lengths. The process of construction the watts-strogatz model has connecting probability p assigned to each node, where ‘p’ varies between 0 and 1 (0-regular, 1-random). While starting from a regular network all the vertices ‘n’ are placed in a topological ring, each vertex is connected to the nearest vertices ‘k’. As shown in example of this procedure where k =4 and n=8. To rearrange the regular network bring randomness into production, each edge connects its end vertex to the end vertex of first-nearest vertex, and it is reconnected to a random vertex with probability ‘p’. The same procedure will be repeated from one vertex to another either in clockwise until one lap is completed. To eradicate the redundancy duplicate/similar/additional edges in the connection process is avoided. Then, the similar process will be continued until the edges connect the vertices to their next nearest neighbours, while every edge tin the regular network has been considered once.

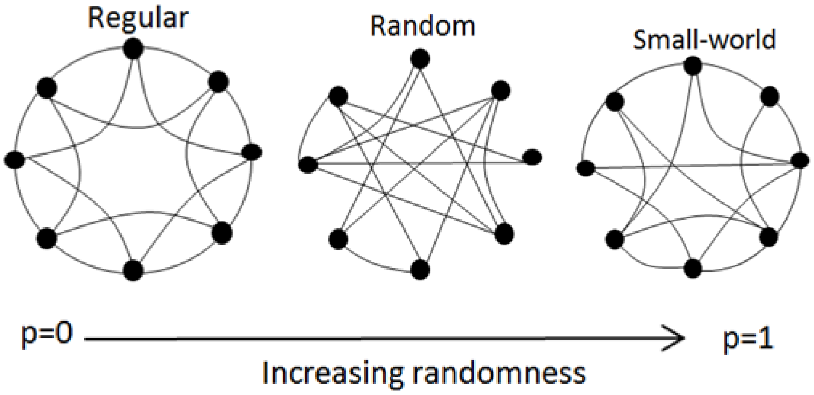


Fig: Watts-Strogatz Model – Reconnecting the Procedure in Random Network.

XI. GENE REGULATORY NETWORKS

While classifying the nucleus, Ribonucleic acid (RNA) which acts as a messenger Ribonucleic acid shortly called as (mRNA) is continuously translated from a portion of deoxy ribonucleic acid (DNA) in response to some intracellular signals and they carried these signals to outer part of the nucleus, there it will be decoded by the ribosomes in cytoplasm for production of proteins and it is responsible for vital functions in a particular cell Deoxyribonucleic acid (DNA) is nothing but a double stranded helix (spiral) molecule which encodes the genetic instructions for each cell functioning and their development. The mRNA is a single stranded molecule which is a small duplicate part of DNA. The different types of proteins like transcription factors controls the concentration of mRNA by blocking or starting the transaction process in the nucleus. The large availability of micro data sets gives a clear information about gene expression changes by time to time. While descriptive network models are constructed using different computational techniques like correlation analysis which exposes the primary knowledge of the regulatory mechanisms.

XII. INTEGRATING SEQUENTIAL GENE EXPRESSION WITH DNA-BINDING DATA

The genome wide DNA-protein binding data, DNA sequence and gene expression data represents, it means translating global and local transcriptional regulatory networks. While combined these different types of data not only improve the statistical power and also deliver complete picture of gene regulation. Yang xie, weipan proposes a novel statistical model to augment protein DNA binding data with gene expression and DNA sequence data when it is available, and also specifies hierarichal bayes model to identify the target genes when compared to conventional approaches on a single data source.

For modeling the gene regulatory networks, essentially we differentiate the physical interactions from association relationships among the genes. While advancing in sequence technology a particular knowledge about thet potential physical binding of number of transaction factors to target genes is available, but the occurrence of the reaction at the time of the transcription commencement is not possible. The sequential gene expression profiles have significant information about biological functions from time –to-time intervals provides more understanding about the possibilities of regulatory interactions behind the time.

However, gene expression profiles integrating with physical binding information of reactions which opens the exact structure of a gene regulatory networks at each time interval. Hence few selected interactions obtained by the profile analysis which may not be exact in case of TF's in the reaction is useless for DNA-binding to target gene and then similar analysis for the reverse condition then gene expression profiles cannot found the specific information about binding data.

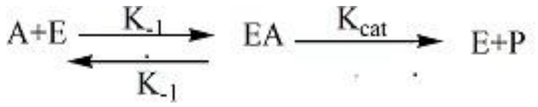
XIII. ESTIMATION OF TIME DELAY IN REGULAR NETWORKS

While transferring the target genes with the effect of transcription factor regulation there is a possibility for the occurrence of time delay. This phenomenon of time delay underlying a gene regulatory networks extracted by using various modeling approaches such as correlation analysis, decision tree related classifiers (soinov, krestyaninaova et.all 2003) and Boolean models (silvescu and honavar 2001). These models are accomplished to approximate delay of one unit time lag, though many regulation processes have multiple unit of time lag during the change of the regulator expression which is transmitted to the corresponding targeted gene. The ability to estimate multiple time delays of gene regulatory will enlarge our data about the development of biological processes.

13.1 Metabollic Networks

The accessibility of whole genomes which permits many new types of analysis and experiments. These new approaches the genomes which make potential and referred to as functional genomics. To define the function of a gene or its protein in the life processes of an organism wherever the function refers and plays its superior role. The chemical system generates needful components like sugar, lipids, amino acids and required components to synthesize them to create a protein and cellular structures. This structure of connected chemical reactions is "Metabolic Networks".

These metabolic networks are together of sequence of biochemical reactions which occurs in a metabolic pathway whereas enzymes catalyze some chemical compounds which produces products for successive interactions of pathway of proteins. These actions are crucial for cell functionality and other sustainability. The simple form of enzymatic reactions as shown below



Enzyme 'E' eases the activation energy of above reaction by required chemical compounds 'A' and forms enzyme substrate compound EA. K_1 association, K_{-1} redissociation and K_{cat} dissociation rates are shown in above equation.

XIV. MODEL GENE REGULATORY NETWORK METHOD

Suppose the gene expression profiles are continuous and linear fashion within each time interval to adopt the linear regression models to find the network structure of a gene protein and estimates the time delay response for each and every pair of target genes and TF's. However, it is necessary to know that regulators have a same target gene it might have different time delays in this model. The steps followed by gene regulatory network model as follows.

STEP 1: Start

STEP 2: Find the transcription factor X_{ij}, interact with each target gene G_j

STEP 3: Time delay t=0

STEP 4: Find the regression coefficient by using LS

STEP 5: $e= (g_j-g_j\text{^})^2$

STEP 6: If time delay is final, end the process.

STPE 7: otherwise check the current delay and update the time delay and start the process again.

--*Multiple-delayed regression model*--

While using a multiple delayed linear regression model to estimate the effect of time delay in TF's on their target genes. Basically, we need to estimate the time delay first and then regression coefficient using the estimated delay. While construct the regression model for a particular gene (assume mice post myocardial infarction (MI)), the TF's and the targeted genes in the particular gene data sets will be identified using TRANSFAC database.

Here genes have been high degree of connectivity in the network it might be potential biomarkers for a myocardial function.

The derived regression model has been evaluated with two statistical tests i.e., adjusted R^2 and ANOVA F-test. The static measures of R^2 measures the percentage of total differences of estimated and actual value of $g_{j.}$ The adjusted R^2 is a modified version of R^2 which includes the number of time points and TF's in its formula.

$$\text{Adjusted } R^2 = 1- (1- R^{2)}\ (N-1)/N-n_{tf}-1$$

Where n_{tf} is the number of transcription factors which interacts with a specific target genes and N is the total number of time points.

As we have different amounts of delay has been estimated for the temporal gene regulatory network. The genes are connected when the regression coefficient related with regulator is not be zero. On each time delay, the network structure is concerning by adding new connections, though we kept previous connections in the network. The fundamental guess in our model is the expression level of target genes in every time point is unnatural by the expression level of the regulatory of earlier time points.

XV. CONCLUSION

The ordinary differential equations and linear regression models are studied for the metabolic and gene regulatory networks of modeling dynamics. For each of the computational techniques reports different features of network dynamics. Temporal expressions of network biological entities could combined with another biological data almost underlying the processes to organize the network dynamics model. The other features of a dynamic process can be estimated the time delay during the modeling process. While evaluating the models statistical measures will be adjusted i.e., Anova F-test.

REFERENCES

[1]. Damian Szklarczyk, Lars Juhl Jensen, Protein-Protein Interaction databases Methods in Molecular Biology Volume 1278, 2015, pp 39-56.

[2]. Smoot ME, Ono K, Ruscheinski J, Wang P-L et al (2011) Cytoscape 2.8: new features for data integration and network visualization. Bioinformatics 27:431–432

[3]. Brown KR, Otasek D, Ali M, McGuffin MJ et al (2009) NAViGaTOR: Network Analysis, Visualization and Graphing Toronto. Bioinformatics 25:3327–3329

[4]. Shaker, Ammar and Lughofer, Edwin. "Self-Adaptive and Local Strategies for a Smooth Treament of Drifts in Data Streams.", Evolving Systems, 5:(4), p. 239-257, 2014.

[5]. Minku and Yao. "DDD: A New Ensemble Approach For Dealing With Concept Drift.", IEEE Transactions on Knowledge and Data Engineering, 24:(4), p. 619-633, 2012.

[6]. P. Bajcsy, J. Han, L. Liu, and J. Yang. Survey of bio-data analysis from data mining perspective. In Jason T. L. Wang, Mohammed J. Zaki, Hannu T. T. Toivonen, and Dennis Shasha, editors, Data Mining in Bioinformatics, pages 9{39. Springer Verlag, 2004.

[7]. Mummery-Widmer, J.L. et al. Nature 458, 987–992 (2009).

[8]. Di Battisa G, Eades P, Tamassia I, Tollis I (1999) Graph Drawing: Algorithms for the visualization of graphs, Prentice Hall.

[9]. Shneiderman, B. (2008). Extreme visualization: squeezing a billion records into a million pixels. In Proc. 33rd ACM Intl. Conf. on Management of Data (SIGMOD), pages 3–12.

[10]. Barabasi, A. L. and R. Albert (1999). "Emergence of scaling in random networks." Science 286(5439): 509-512

[11]. Hintze, A. and C. Adami (2008). "Evolution of complex modular biological networks." PLoS computational biology 4(2): e23.

[12]. Jeong, H., et al. (2000). "The large-scale organization of metabolic networks." Nature 407(6804): 651-654.

[13]. Li, X., et al. (2006). "Discovery of Time-Delayed Gene Regulatory Networks based on temporal gene expression profiling." BMC bioinformatics 7: 26.

[14]. Li, Z. and C. Chan (2004). "Extracting novel information from gene expression data." Trends in Biotechnology 22(8): 381-383.

[15]. Mohyedinbonab, E., et al. (2013). Time delay estimation in gene regulatory networks. System of Systems Engineering (SoSE), 2013 8th International Conference on.

[16]. Yu, H. and M. Gerstein (2006). "Genomic analysis of the hierarchical structure of regulatory networks." Proceedings of the National Academy of Sciences 103(40): 14724-14731.