Study Biological Networks For Application of machines learning Techniques.

Ruchika Sharma Gaurav Kumar Jain **Ravi Ranjan****

Scientists, biologists and medical practitioners are always enthusiastic to know the human unconscious control of body activity. A simple question of: how the cells talk? Can answer several mysteries of cognitive science and life control of species .As the nature manages and organize an organism' s perception, circulation, respiration, digestion ,excretion and action, if we can ably simulate the language of cells, consequently establishing the associated bimolecular interaction networks, it can yield miraculous results procuring digital control on our body with a profound impact on biotechnological and pharmaceutical fields studying the properties of protein and gene interaction network promises to yield prediction of cell language which can shed now light on the evolution of species and also improve our understanding of the underlying biological phenomena. Science a large part of these complex biological networks is governed by probabilistic theory and undefined rules that comes from the incompleteness of detailed knowledge on system biology, the modern machine learning algorithms using probability and heuristics are the major techniques for their simulation. Several distinct and complementary learning algorithms are being explored to infer networks in biology. In this paper, we highlight application of various machine learning techniques to study biological networks inferring organized complexity of the languages of cells and providing a context for reader new to the field.

Keywords: Machines leaning, systems biology, biological database, biological network, graph model, support vector machine, natural network, classification trees, genetic algorithm

Introduction

Machine learning to refer to a system capable of a autonomous acquisition and integration of knowledge, thus programming computer to optimize a performance criterion by using sample data or post experience. This capacity to learn from experience and analytical observation results in an adaptable system that can continuously self-improve and thereby offer increased efficiency and effectiveness (Mitchell, 1997).the optimized criterion can be the accuracy provided by a predictive model in a modeling problem. In a modeling problem, the term 'leaning' refers to execution of a programme tom induce a model by employing training data or past experience. Machine learning uses statistical theory when building computational models since the objective is to make inference from a sample. Two major step in this process are to induce the model by processing a large amount of data and to represent the model making inference efficiency. In an optimization problem, the task is to find out an optimal solution in a space of multiple (sometimes exponentially sized) possible solution. The choice of the optimization method to be used is crucial for the problem solution.

Machine learning research has been making great progress in many directions. Once of them does the field of computational system biology comprise three prominent branches of science: biology, computer science and complexity theory. Systems biology is emerging as a new challenging field for scientists. Complexity in natural science and system biology typically arises in biological system with large number of components and a large number of structured interactions between them. These structured or patterned interactions in multi component system make accurate prediction of system behavior from simulation very difficult.

Weaver was first to define 'disorganised compledxity' as a problem, in which the number of variable is very large and of these variables is best described as a random process (weaver, 1948).This was at the 'molecular level' and the most successful simulation method for representing phenomena at this level derive from statistical consideration. In the context o the cell, matters are complicated by the fact that organization becomes an essential feature of the process under consideration. Hence at the 'cellular level', a large number of interrelated factors are integrated, working as a single complete system,

which in general, cannot be dealt in the Newtonian realm of physics or mere statistical modeling. Weaver, at that, addressed this problem of 'organized complexity' as a challenge for science in the coming 50years.

Followed by weaver, Ludwig von bertanffy's system theory (bertalanffy, 1996),haken's synergetics, .chaos theory, the science of self organized criticality (Kauffman,1995;and bak, 1997), non-equilibrium physics, power laws, and the availability of modern experimental techniques (hi-tech microscopy, laser tweezers, nanotechnology, DNA microarrays and mass spectrometer) with generation of vast amount of biological data renewed interest in complexity studies and system biology in the last decade or so. While the technology to generate and Manage data races ahead, it become apparent that mythological advances in the analysis of data are urgently required if we want to turn data into knowledge. Since the organized complexity of cells follows probabilistic derivations with a NP-hard biological decision-making problem, the self adaptable machine learning algorithms are proving to be a paramount simulation technique for them.

2. System Biology: The Study of Complex Networks

One of the most difficult and central topics in systems biology is how be systemic properties of cells and organisms from data, model them, and take into account in data analysis. It has been widely appreciated that most of the earlier biological research has focused on studying only parts of the system (DNA,protins and genes) and understanding how these biological parts interact is still a next big challenge.The study of biological network hold tremendous promise for a number of aspects of the drug discovery and development process. Impact are possible in diverse areas like biotechnology , bioinformatics or large-scale genomic data analysis. A hallmark of post-genomics is the development of high-throughput method for the analysis of complex biological system. In consequence , it is increasingly commonplace to have access to large database of variables ('omits data proteomics , genomics)against which research into the collection , modeling and analysis of biological is not yet mature, and there is still much to be done in order to bring capability in the integration , manipulation and analysis of such complex network to maturity. The modern machine learning methods comprising automatic reasoning tools and exhaustive search procedure allow for flexible data integration, inherent inferevcing capability and efficient modeling ability for such complex network networks.

2.1 Computational Challenges

Modeling and analysis of complex biological networks has spurred increasing interest in the field of computational biological and the biostatistics communities. Biological need rigorous and flexible tools to describe, infer and study these complex system, For a long time quantitative mathematical models have investigated the dynamic of bimolecular system by developing numerical models involving (highly nonlinear)differential equations. It provided a firm ground for the numerical analysis of biological systems. However, these quantitative models can hardly be reused and composed with other models in a systemic fashion, and are limited to a few tenths of variables (chen et al.,2005).

A major computational challenge in system biological consists in identified with reasonable accuracy those complex macromolecular interactions that take place at different level from genes to metabolites through proteins. Once identified, a network model can be used to simulate the process it represent, for a variety of analysis, ranging from statistical properties of its topology to predictions of feature of its dynamic behavier, or even prediction of cellular phenotypes. Another challenge is the modularity and compositionality of biological models. It is not an easy task today to combine given models of different pathway sharring some molecular component in a given organism, and obtain a mixed model of the complex system. This is a restriction to the reuse of models and to their direct use in any application. A third challenge for system biology is to go beyond simulations and use models to automate various forms of biological

reasoning in purely qualitative models too. Computer aided inference of interact or network , or computer aided drug target discovery, need nontrivial automatic reasoning tools to assist the biologists.

A summary of simulation requirements posing computational challenge for complex networks in system biology (kitano, 2002) can thus be listed as follows:

. Methodologies for parameter selection.
. Identification of causal relationship, feedback and circularity from experiment data.
. Investigation into the stability and robustness of cellular systems.
. Modular representation and simulation of large scale dynamic systems.
. visualization and fusion of information, integration of models and simulators.
. Experimental and formal methods for model validation .
. Scaling models across scales and description levels (from genes and protein to organisms).

2.2 Network Database: The Biological Database

Rapid advances in high throughput technologies and large scale experiments in biology are providing as with breathtaking new insights into cellular machinery and processes. The public availability in interaction network database containing thousands of interactions for a number of species is a highlights of these advances. There are many available database of biological interaction data which can be used as training datasets for learning algorithm and in various other simulations models. Many properties of these networks have already been studied and these studies have yielded many important results in the following sections, we describe various biomolecular interaction database that are currently available.

2.2.1 BIND: Bio molecular Interaction Network Database

The BIND (Gilbert, 2005) is a collection of records documenting molecular interactions. BIND contains interaction data for a variety of organisms like mouse, yeast, HIV virus, etc. the interaction data has been obtained from high through experimentally verified data submissions and handwritten from scientific literature.

2.2.2 DIP: Database of Interacting Proteins

The DIP (Xenarios et al., 2000) contains experimentally determined interactions between proteins for a large number of organisms like human, yeast, mouse, worm, etc. the interaction in the DIP database have been generated by combining the information from a variety of sources. The stored was collected manually as well as using computational approaches.

2.2.3 GRID: the general repository for interaction datasets

The GRID (breitkreutz et al., 2003) are available foor yeast, fly and worm. GRID contains physical, genetic and functional interactions between proteins. The data has been generated from biological experiments such as the two hybrid system, affinity precipitation and synthetic lethality. Yeast network has 4,920 vertices and 17,816 edges. Fly network has 7,940 vertices and 25,665 edges. Worm network has 2,803 vertices and edges.

2.2.4 KEGG pathway database

The Kyoto encyclopedia of gens and genomes (KEGG) (Ogata et al. , 1999) is a pathway database from the kanehisa laboratory of Kyoto university bioinformatics center, to understand systematic function of the cell or the organism from its genomic information. It is one of the major repositories of biological networks. It has a standard file format, KGML, TO distribute biological network information. KGML defines objects of the biological network and their relationships as an XML data base structure.

2.3 Aspects of machine learning

Interactive systems in biology mainly consists of DNA, genes, proteins, metabolites, inhibitors and cofactors. The relations include biochemical

reactions, in which one set of biological component is transformed to another in a biochemical reaction catalyzed by an enzyme. These complex systems can be described as many particle system with, however, non identical particles, such that many traditional approaches from statistical mechanics and condensed matter theory is not applicable. Moreover, the experimental data usually have high noise levels and are under sampled. Traditional statistion modeling here reaches a point, where it can given a very concise mathematical description of the fundamental phenomenon governing the system, but is not able to give accurate predictions for outcomes of future experiments. For this task, other methods are needed like self adaptive tools (the machine learning algorithms), which can learn predictive models from under sampled and possibly noisy data.

Modeling frameworks for biological networks, with the existing methods can identify models from data within these frameworks. Once a formal framework is defined to describe models of biological networks, the question problem can be expressed in the framework of machine learning. Given a family of mathematical models of gene interactions and asset of observations, learning consists here in optimizing the parameters of the model in such way that it captures the observed behavior of the true system. The ability of the instantiated modal to be used in prediction is referred as the generalization property. A model is able to generalize if learning ensures a tradeoff between a good fit to the data and simplicity of the models .solving a learning problem leads to three key question: the representation problem, the optimization problem and the validation problem.

The representation problem concerns mostly the choice of the formalism, in which data and the model are going to be expressed, and the method of encode them into this formalism , both symbolic and numerical leaning lead to an optimization problem whose nature is combinatorial approaches are solves using heuristics to ensure a large exploration of the models spaces. At last, validation is required to identify how one can trust the inferred model (Florence and schachter, 2006).

Some of the intrinsic interest in the system biology area from a machine learning perspective includes:

The availability of large scale background knowledge on existing known biochemical networks from publicly available resources, such as KEGG, DIP.GRID, BIND (SECTION 2.2 ABOVE);

An abundance of training and test data from a variety of sources including microarray experiments and metabolomic data from NMR and mass spectroscopy (wet) experiments;

The inherent importance of the problem owing to its application in biology and medicine; and The inherent relational structure in the form of spatial and temporal interactions of the molecules involved.

Machine learning models, thus have wide

Potential application in systems biology. For instance, in the new area of personalized medicine techniques which allow the construction of models of the toxic reactions of individuals to drug treatment would be of frat benefit. There are several reference books on machine learning topics (Mitchell, 1997; and hastie et al. 2001). Recently, some interesting books intersecting machine learning and computational biology domains have been published (Durbin et al., 1998; pevzner, 2000; baldi and brunak, 2001; forgel and corne, 2002; frasconi and Shamir, 2003; seiffert et al., 2005; mitra et al., 2008; and lodhi and muggleton, 2010).

3. Machine learning applications to study biological networks

The qualitative machine learning models of bimolecular interactions nowadays constitute the core of cell systems biology. Interaction diagrams are the first tool used by biologists to reason about complex systems. The accumulation of knowledge on gene interaction and pathways is currently entered in databases, such as KEGG. The KEGG (ogata et al., 1999), in the form of annotated

diagrams. Tools, such as biospice, gepasi, e-cell, etc., have been developed for making simulations based on these databases when numerical data is present. Furthermore, the interoperability between databases and simulation tools is now possible with standard exchange formats such as the system biology markup language (SBML) (hucka et al., 2003) and system biology workbench (hucka et al., 2001) allowing large scale simulations and integration of models with simulation tools respectively.

Since machine learning techniques are essentially data driven usually large amount of data are needed learning algorithm to be applicable they play a key role in advancing systems biology. The most used approach in system biology is the probabilistic graphical models. The advantage of using graphical paradigm to model biological networks are (1) they are based on probability theory, a scientific discipline with sound mathematical development; (2) probability theory could be used as a framework to deal with the uncertainty and noise underlying biological domains; and (3) inference algorithms (exact and approximate) developed in these models enable different types of reasoning inside the model . Larrinaga et al. (2005) successfully applied probabilistic graphical models in the gene regulatory network. a novel application of new probabilistic graphical models to infer the generic networks. A novel application of new probabilistic graphical models in the gene regulatory network was shown by wang et al.(2005). Roy et al. (2009) proposed a novel approach for scalable learning of large networks, called as cluster and infer networks (cin).

Support vector machine (svms) were sufficiently utilized by minakuch et al. (2002) to develop a reliable prediction system of protein interaction sites on the protein surface from their three dimensional structure. Supper et al. (2006) presented a critical evaluation of the application of various machine learning techniques, viz., multiple linear regression, SVMs, decision trees and Bayesian network to infer gene regulatory network. The performance of these methods is assessed by comparing the topology of the reconstructed models to a valdation network. A recent work of bui et al. (2010) used SVMs to classify features specific candidate PPI pairs. Their system achieved the best performance on cross corpora evaluation and comparative performance in terms of computational efficiency.

A neural network based algorithm efficiently implemented by fariselli et al. (2002), increased the predictive performance of protein-protein interaction sites in protein structures. The goal was to reduce the number of spurious assignment and developing knowledge- based computational approach to focus on clusters of predicted residues on the protein surface (fariselliet al., 2003). Keedwell et al. (2002) utilized artificial neural networks to construct gene regulatory networks. Mostafavi et al. (2006) applied neural networks to model regulatory interactions in temporal gene expression data. The trained neural networks predict the expression profile of a gene, utilizing a minimal set of input gene profiles, with high accuracy on the test data. Eom and zhang (2006) proposed a method for prediction of prediction of protein interaction with neural network-based feature associate rule mining. Chen and liu (2006) proposed forward pruning decision trees and neural network for domain-based predictive model of protein-protein interaction networks (ANNs) were explored by mitra et al.(2009) targeting the computation efficiency in order to generate genetic networks.

Several other works have shown the application of Bayesian networks to model genetic networks (friedman et al., 2000; hwang et al., 2001; chang et al., 2002; and lee and lee, 2005) and protein interaction networks. In tamda work, DNA sequence in formation is mixed with microarray data in the Bayesian network in order data is limited (tamada et al., 2003). Nariai et al.(2004) estimated genetic networks from expression data being refined using protein-protein interactions. Husmeier (2003) tested the viability of the Bayesian network paradigm for gene network modeling. These dynamic Bayesian networks were able to show how genes regulate each other across time in the complex workings of regulatory pathways. Different worls have

considered the use of dynamic Bayesian networks to infer regulatory pathways (Murphy and mian, 1999; ong and page, 2001; and sugimoto and iba, 2004).

Classification trees were used for modeling signal-response cascades, and the methodology was applied to predict the call migration aped using phosphorylation levels of signaling proteins (hautaniemi et al., 2005). Steffen et al. (2002) used clustering methods applied to microarray data and protein-protein interaction data are combined in construction of a signal transduction network. Middendorf et al. (2004) used boosting with classification trees as base classifier for the prediction of a gene regulatory response, which is considered a binary variable.

There are various applications of Genetic Programming (GP) to the inference of gene networks (Sakamoto and Iba, 2001; and shin et al., 2002). Earlier, GP has been applied to select regulatory structures (Gilman and Ross, 1995)and estimate the parameter of bioprocesses(park et al.,1997). The identification of transcription factor binding sites has been treated using markov chain optimization (Kyle et al., 2002). G has also been applied to model genetic networks (shinichi et al., 2003). The inference of genetic networks has been achieved using many other evolutionary algorithms till date (Kimura et al., 2005;Noman and Iba, 2005; and Sirbu et al., 2010).

Other example include application of Genetic Algorithms(GA)to infer a biological network. For instance, Shin and Iba (201) have shown an application of genetic algorithms to the gene network inference problem. The GA is applied to train the model with observed data to predict the regulatory pathways, represented as influence matrix. This approach can be applied with small amount of data, optimize large amount of parameters simultaneously, and can be applied on nonlinear models. GA implementation include multiple stage evolution and matrix chromosomes. Iba and Mimura (2002) has also shown evolutionary computing to infer GRN. They establish the system that realizes their approach by GA based interactive algorithm. Experimental results showed that method proposed can infer the network structure accurately with a relatively small amount of expression data.

Yet another optimization algorithm based on parallelized GA for inference of biological scale free network was proposed by Tominaga et al.,(2003). The optimization task was to infer a structure of biochemical network only from time series data of each biochemical element. This is a inverse problem which cannot be solved analytically, and only heuristic searches, such as GA simulated annealing, etc. are practically effective. The authors claimed that result showed high parallelization efficiency of proposed GA based algorithm. A hybrid approach by ressom et al., (2006) presented a novel algorithm that combines a Recurrent Neural Networks (RNN) and two swarm intelligence (SI) methods to infer a gene Regulatory Network (GRN)from time course gene expression data . the algorithm sues Ant Colony Optimization (ACO) to identify the optimal architecture of an RNN, while the weights of the RNN are Optimized using Particle Swam Optimization (PSO). The proposed hybrid SI RNN algorithm to infer networks of interaction from simulated and real world gene expression data allows to gain new insights into the field of machine learning.

Conclusion

Nowadays one of the most challenging problems in computational systems biology is to transform the house volume of biological data, provided by newly developed technologies into knowledge . machine learning has become an important tool to carry out this transformation. The self adaptable machine learning algorithm are able to model framework for large biological networks giving scope to a variety of structural analysis and predictive tasks and can efficiently represent the complex information which is associated with such networks-be it information about the physical, structural or functional properties of the biological nodes, the topology of the network, or the uncertainty associated with intermolecular reactions. The continues effort of scientists and researchers all over the world may

improve these learning techniques incrementally as new constraints are added with time, ensuring some robustness in computational procedures with respect to modifications, while reserving simplicity and tractability.

In this paper, we have described various machine learning approaches to model and infer the interactive networks of biomolecules and biological components. The paper also highlights the importance of biological networks the talking language of cells, in life control of species and justifies the applicability of machine learning approach towards the solution of a NP Hard complex network inference problemwhich is largely governed by uncertainty principal. A brief introduction to computational systems biological database freely available on the internet is also given for readers new to the fields.

Finally, it can be concluded that modeling with in system biology is a key application area for machine learning in general. The studies described in this paper indicate that learning algorithms have the potential to be a key technology in the area which is now drawing mezor scientific interest internationally. The paper can serve as a gateway to some of the most representative works in the field and as an insightful categorization and classification of the machine learning methods in systems biology.

References

1. Bak p (1997), how Nature Works the Science of Self Organized Criticality, p.21 Oxford University press, Oxford, UK.
2. Baled P and Brunk S (2001), Boinfornanties, The Machine Learning Approach p. 452, The MIT press.
3. Bertalanffy L von (1969), General Systems Theory: Foundations, Development Applications, Pub-id 101-279-836, George Braziller, New York, USA.
4. Breitkreutz B J,Strark Cand Tyers M (2003), "The GRID: The General Repositor
   For Interaction Datasets", Genome Biol, Vol.4, No.3, R23.
5. Bui Quoc- Chinh, Katrenko Sophia and sloot Peter M A (2010), "A Hybrid Approach to Extract Protein-protein Interactions", Bioinformatics, Vol. 27 No.2, pp.259-265.
6. Chang J-H, Hwang K-B and Zhang B T (2002),"Analysis of Gene Expression Profile And Drug Activity Patterns by Clustering and Bayesian Network Learning", Methods of Microarray Data Analysis' II,pp.169-184,Kluwer Academic Publishers.
7. Chen Xue-Wen and Liu Mei (2006),"Domain Based Predictive Models for Protein
   Protein Interaction Prediction", FURASIP, J. on Applied Signal Processing pp.1-8.
8. Chen K C, Calzone L, Csikasz-N a et al (2005). "Integrative Analysis of Cell Cycle Control in Budding Yeast", Molecular Biology of Cell, Vol. 15, pp.3841-3862.
9. Durbin R, Eddy SR, Krogh A and Mitchison (1998), Biological Sequence Analysis Probabilistic Models of Proteins and Nuclear Acids, p. 350, The Cambridge University press.
10. Eom jae-Hong and Zhang Byoung-Tak (2006), Prediction of Protein Interaction with Neural Network-Based Feature Association Rude Mining,pp.30-39,ICONIP'06, Part III, Springer Verlag Berlin Heidelberg Lecture Notes in Computer Science (LNCS) 4234.
11. Fariselli P, Pazos F, Valencia A and Casadto R (2002), "Prediction of Protein Protein Interaction Sites in Heterocomplexes with Neural Network", Eur, j.Biochen. Vol.269, No.5, pp.1356-1361.
12. Fariselli P, Zauli A, Rossi I et al. (2003),"A Neural Network Method to Improve Prediction of Protein-protein Interaction sites in Heterocomplexes", Neural Network for Signal Processing (NNSI') apos03, IEEE 13th Workshop, 17-19th September 17-19, 2003.