Challenges in An Adaptive Modeling Framework for Systems Biology

Rohit Joshi*, Tze Yun Leong *

* Medical Computing Laboratory,
School of Computing,
National University of Singapore,
3, Science Drive 2, Singapore 117543
{rjoshi, leongty}@comp.nus.edu.sg

Abstract

With advances in biology and medicine, there is a need for new decision support systems that can integrate the knowledge of these domains and enhance the decision making process. Several issues need to be addressed before we can design an intelligent biomedical decision support system. With rapid speed of development and innovation, biomedical information is continuously changing, so systems adaptive to change in knowledge are needed. Furthermore, successful integration of knowledge from experimental data as well as that stored in textual databases is needed. In this paper, we discuss some of the challenges in an adaptive modeling framework for complex systems. We focus on systems biology and discuss the challenges in two aspects – modeling from experimental data and modeling from scientific text articles. Firstly, we focus on learning from experimental data and address why adaptive behaviour is required. Secondly, we discuss the importance of having a general adaptive system that may be able to extract knowledge from text for several domains rather than one specific domain as is done in most of the current state-of-the art systems.

Introduction

Decision Support Systems are designed to improve the processes and outcomes of decision making. Several successful systems based on medical knowledge have been deployed for medical decision making. With advances in biology and medicine, it will soon become important to make decisions based on both biological and medicinal knowledge. Furthermore, scientists have initiated efforts to understand the collective behaviour of the body systems and explain the observed phenotypic behaviour by integrating knowledge at genomic, proteomic, metabolomic and phenotypic levels. Hence, there is soon going to be a need for intelligent biomedical decision support systems. However, several issues need to be addressed before we can design an intelligent biomedical decision support system.

With rapid speed of innovation and advancement in biomedical field,, biomedical information is continuously changing, so systems adaptive to change in knowledge are needed. Furthermore, successful integration of knowledge from experimental data as well as that stored in textual databases is needed. In this paper, we discuss some of the challenges in adaptive modeling framework for complex systems. We focus on systems biology and discuss challenges in two aspects – modeling from experimental data and modeling from scientific text articles.

Firstly, we talk about the challenges in modeling complex systems of an organism from experimental data. Scientists have initiated efforts to understand the collective behaviour of the systems of an organism. Numerous biological and medical databases have already been built to provide comprehensive biomedical information. Scientists also have abilities to measure the expression levels of thousands of cellular entities simultaneously; however maps of interaction and regulation networks needed to understand the collective behaviour are still not completely known. Learning of these interaction maps is essential before we can design an intelligent biomedical decision support systems. However, learning from the data may not be enough and it is essential to have an adaptive component in these learnt networks as the biological networks behave differently to different inputs.

Secondly, we focus on the challenges in modeling complex systems from scientific text articles. Currently, researchers have been focusing on building systems for specific domains. For using the system in a new domain, it is required to either change the underlying system or retrain it. Retraining requires a lot of new training examples; hence leading to wastage of time and effort. Furthermore, to understand the collective behaviour, it is needed to integrate knowledge from diverse domains. Hence, systems that are adaptive to diverse domains are required. We discuss our efforts and ideas in building such a system.

Challenges in modeling complex systems of an organism from experimental data

This section focuses on some of the challenges that need to be addressed before an intelligent decision support system

may be designed for complex systems. The study of complex systems is about studying how parts of systems may interact with each other and give rise to the collective behaviour of the system. Intelligent decision support systems are generally designed to improve decision making when the information is incomplete or uncertain. We have to address challenges to some of the core components of the decision support system such as the knowledge base and the inference mechanism. Knowledge base of a complex system may have information about the names, location, functions, interaction maps etc. of the different parts of the systems. However, in the real world, this information is generally incomplete and uncertain. For inference mechanism, we need to be able to do inference at multiple levels of hierarchy in a complex system.

Our focus is on the emerging field of Systems Biology [Kitano, 2002], which is about studying complex systems of an organism. Systems Biology field has emerged following the great advances made in the biomedical field in the past decade. Numerous biological databases such as Gene Ontology, SwissProt, KEGG, PathDB have been built to provide catalogue information about the names of genes, proteins, metabolites and their functions, structures and localizations. These knowledge bases are only a comprehensive part-list collection of the organisms and are not enough to explain the collective behaviour of the body systems such as the observed phenotypic behaviour.

Recent advances in high throughput analysis have enabled scientists to measure expressions of thousands of cellular entities simultaneously. This has opened a way for scientists to try and model the networks of interactions and regulations at the cellular level first. Interaction in a cell form three different networks – gene networks, signal transduction networks and the metabolic networks. Modeling these networks may reveal the concise picture of interactions, their strengths and possibly the emerging behaviour of the system and will provide better insights into the observed phenotypic behaviour. However, modeling these networks from experimental data has presented a significant challenge to the researchers as explained below.

First, one of the crucial missing components in the current knowledge bases before an inference mechanism can be applied to even the smallest of the organism is a complete interaction and regulation map of the body networks. Secondly, we need to understand and incorporate the underlying dynamics of the systems in the knowledge base and inference mechanism. Thirdly, we need a modular inference mechanism that can infer across multiple levels of hierarchy in the body systems. Fourth, biological networks react differently to change in environment, perturbations (inputs), constraints and time, so the framework needs an adaptive component that can adapt the learnt network accordingly to these changes. In this section, we concentrate only on learning the interaction and regulation map from data because this itself is extremely challenging.

Learning and modeling these biological networks is challenging for a number of reasons: First, the curse of dimensionality in handling hundreds and thousands of variables (genes, proteins or cells) makes it computationally challenging. Second, the data is scarce. Only few sample points are collected due to costly experiments. Modeling in high dimension usually requires large sample datasets. Third, experimental data is usually noisy and incomplete. Fourth, biological networks are non-linear dynamic systems with feedback, which is quite challenging to model. Also, there are some current limitations such as experimental limitations to what we can measure. Moreover, biological networks are complex systems and require synergetic integration of heterogeneous datasets at multiple levels of hierarchy in an organism. Finally, learning may need to be adaptive to change in environment, perturbations or constraints.

Literature Review for modeling cellular networks

Several techniques have been proposed to model gene networks, protein interaction network and metabolite network. However, none of the previous works have focused on adaptive behaviour.

Differential equation modeling has been the most popular [Chen et.al, 1999] because it can model at finer levels of detail. However, differential equation modeling technique is not scalable to higher dimensions. Boolean networks have also been used to model the cellular networks [Akutsu et al., 1999]. However, Boolean networks theory has quite strong simplifying assumptions that enable large regulatory networks to be analysed easily. Linear and non linear auto-regression models [D'Haeseleer et al., 1999] have also been used to describe the relationships between the regulatory gene and regulated gene in linear or non linear regression coefficients. Murphy [Murphy et al., 1999] has shown that both boolean and linear and nonlinear auto regression models can be considered as the special cases of the dynamic Bayesian networks.

Recently few researchers have tried to apply Bayesian network(BN) to model gene networks [Friedman et al., 2000]; [Ong et.al, 2002]; [Perrin et.al, 2003]; [Rangel et al. ,2004] [Li et al, 2005]. Existing approaches can be divided along the following dimensions:

a) Static Vs Dynamic BNs: Static BN approaches [Friedman et al., 2000] cannot model the cyclic feedback, a common feature in biological systems, where as Dynamic BNs approaches [Ong et al., 2002]; [Perrien et al., 2003] can handle cyclic property by unrolling networks over time. But this has serious implications on the computational cost.

b) Discrete Vs Continuous data: Several authors [Friedman et al., 2000]; [Ong et. Al, 2002] have used discrete data. Discretizing data looses important biological information.

c) Handling Hidden Variables: [Rangel et al., 2004]: Handling hidden variables is a desired property.

d) Exact Vs Approximate Learning: With high dimensions and hidden variables, exact methods can become intractable. Approximate methods may have an advantage.

e) Fully observed Vs Partially Observed data: It is difficult to have a fully observed dataset. Most of the high throughput datasets are partially observed. However, sometimes missing values can be approximated using statistical techniques such as mean, median or some other sophisticated way.

f) Variable Reduction Method: Li [Li et al, 2005] has proposed a framework for learning Bayesian network using variable grouping. The learnt network gives a higher level abstraction of biological functions in a cell and is robust enough to handle different experimental datasets.

Proposed Methodology for modeling cellular networks

We argue that Probabilistic Graphical Models (PGMs) are powerful techniques and have several advantages for modeling cellular networks. PGMs can not only model a coarse-grained view of cellular machinery but also incorporate much finer level of details such as transcription factor binding rates, which may make modeling more realistic. PGMs are stochastic in nature, known to handle noisy and missing data and incorporate latent effects from unmeasured quantities. Furthermore, PGMs can be applied to model linear as well as non-linear dynamics of the systems. PGMs can provide most probable explanations and also answer 'what-if' scenarios. PGMs are also adaptive, unlike many other techniques, and can incorporate new knowledge. Biological knowledge can also be integrated in the form of priors.

The main challenge in learning DBNs is that there are hundreds of variables and only few sample points. Learning DBN procedure involves model selection. The usual practice for model selection is to pick the highest scoring model. Then, use that model to infer the domain structure. Small size of samples or data scarcity lead to many high scoring models and the posterior probability over structures, P(M|D) is diffused. So, answers based on single model selection theory are often incorrect. It is more appropriate to select features that are common to many models.

Previous works using Probabilistic Graphical Models have focused on construction of network topology. Adaptive learning is yet to be addressed. Network topology is the concise picture of interactions of genes and proteins e.g., Figure 1 is a network topology of genetic network of galactose utilization mechanism. Currently, research works have aimed at learning the complete structure from experimental data. It means that they try to learn all possible interactions in the genome from the data. In the example of galactose utilization problem, they will try to learn complete Figure 1 from few sample points. Although this may appear reasonable at first, the current algorithms identify only few known correct edges. Many more edges, for which we are not sure if they are wrong or right, are also usually found by these algorithms. Probably a lot of them are wrong.

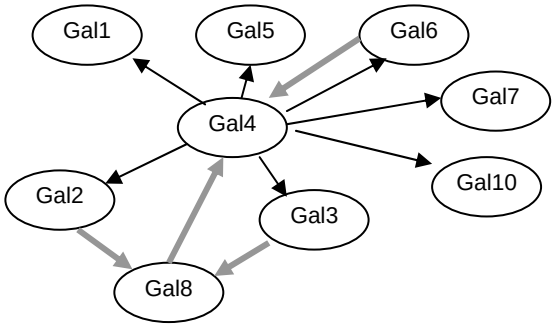

Figure 1: Genetic Network for Galactose Utilization

Gal1 to Gal10 are genes involved in this pathway
Black Arrow: Activates (+ve influence);
Thick Grey Arrow: Inhibits (-ve influence)

We conducted several experiments to learn dynamic Bayesian networks structure at the global scale from data. Due to data scarcity, the learnt models usually contain many false positive interactions and only few true positive interactions. Our observation is that although it may not be useful to learn at the global scale, but it may be more meaningful to learn the small local structures- which we call the sub-networks. Sub-networks are regions of the network that show significant changes in expression over particular sets of conditions. We would also like to point out that learning the sub-networks from gene expression, protein expression or metabolite concentrations may act as building blocks for constructing larger network by combining them together in a modular fashion.

We are currently working on an algorithm based on Markov chain Monte Carlo(MCMC) approximate learning methods to discover sub-networks with high confidence. Approximate learning methods may have an advantage over exact methods as exact methods can not be used if the learning problem becomes intractable. Since biological networks involve a large number of variables and some of them may be hidden, learning problem may become

intractable. Using MCMC approach, it is easy to incorporate priors in the form of known connections. Also, constraints such as sign of interaction (+ve or –ve) can also be used to study different choices for the priors. However, MCMC methods are known to be slow, and it is usually difficult to assess convergence.

Future Work

Biological networks react differently to change in environment, perturbations (inputs), and constraints. We are currently thinking of ways to make the system adaptive so that it can adjust the interactions between the sub-networks according to the perturbations applied to the network.

Challenges for modeling complex systems from scientific text articles

In biomedical domain, scientists need to have the latest knowledge for decision making. MEDLINE database generally has all the latest biomedical knowledge published in the scientific articles. Researchers in information retrieval and information extraction fields have devoted a huge effort to mine the knowledge from the MEDLINE database. However, they have normally been specific to a domain such as radiology or pathology. These systems either use pre-defined regular expression patterns or employ an inductive classifier learnt using lots of training examples. Providing good training examples requires domain experts to spend a huge effort and time. Regular expression based approach is generally very restrictive. The challenge comes when we require a system that can be used in multiple domains. In general, such a task will require multiple systems to be deployed with each specializing in one particular domain. Furthermore, to understand the collective behaviour of the systems, it is needed to integrate the knowledge form diverse domains. Hence, systems that are adaptive to diverse domains are required.

Another challenge is to find relationships among this extracted information. Relationships reflect interactions among different concepts. Building such relationships structure is like extracting a Bayesian network from the text but without the numerical probabilities. Generally, the relevant structure in biomedical decision making task is drawn by the domain experts. So, some scientists may wonder why we want to extract structure from the text. However, in domains such as Systems Biology where the detailed knowledge is not yet known and is continuously being updated at a rapid rate, it becomes important for a decision support system to have an automatic agent to update its knowledge base with the relevant structure. In systems biology research, the scientists want to understand the networks of interactions of disease, metabolites, proteins and genes. Extracting just the relevant structure of interactions from scientific articles has thus become extremely important.

Proposed Methodology

This section focuses on two different challenges: First, to have a system that can adapt itself to diverse domains; Second, to construct relationships structure among the concepts from scientific articles. We are currently working on one such adaptive system using a combination of biomedical ontology and data mining techniques with Natural Language Processing (NLP). Idea is to use minimum training examples to train the classifier and have a self learning and improvement quality. For extracting relationships, we use biomedical ontology to map the sentences to the semantic concepts and semantic types. By training the system on the semantic concepts and types rather than the keywords, we significantly decrease the amount of training data required to learn the structure.

We have proposed a semi-supervised Mutually Reinforcing Learning (MRL) algorithm [Li, 2004] to classify biomedical citations without the need of any manually labeled training data. Our system automatically labels the data needed for training for different tasks. We use keywords based search engines to retrieve relevant documents. These retrieved documents are weakly labeled positive data. Usually, a classifier trained on such a training data will be a weak classifier because of the noise present in the training examples. This restricts the usefulness of such approaches for training the classifiers. However, our proposed semi-supervised self-learning algorithm iteratively obtains a clean training set needed for the classification task. The algorithm adds the semantic concepts and types from the biomedical ontology to the training dataset and iteratively retrains the classifier till a clean training dataset is obtained. Our preliminary results showed that clean training set is obtained from the raw original data set. We have used it successfully for classifying various biomedical citations.

For extracting relationships from text, our system [Joshi et al., 2004] again uses the biomedical ontology and maps the sentences in the document text to its semantic concepts and semantic types. We use association rule mining to find the significant relationships among the concepts rather than keywords in the text. This reduces the number of relationships found but increases the quality of the relationships found. We argue that the choice of features used to represent a domain has a profound effect on the quality of model produced. Association rule mining (ARM) produces much better results when semantic concepts are used than the keywords. However, relationships extracted using ARM still need to be verified. We are trying to build a relationship verification system to verify these relationships found. In our verification

module, we are experimenting with relational learning functions and Natural Language Processing (NLP) techniques.

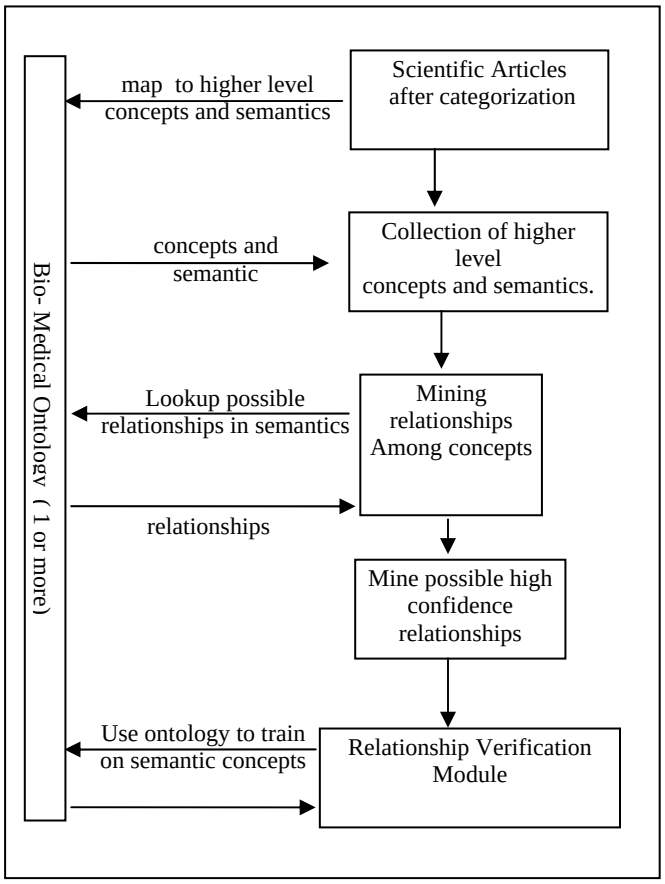

Figure 2: Architecture of our system for relationship extraction

The proposed architecture is dependent on successful combination of biomedical ontology databases with the data mining and natural language processing technique. We believe that the general and adaptive systems are quite possible to be constructed with intelligent use of advances in different fields. In future, we would like to use our architecture to extract gene, protein and metabolite networks from the text.

Literature Review

In recent years, researchers have investigated the idea of reducing the manual labeling effort, and have proposed semi-supervised learning techniques. These are of mainly two types: 1) Labeled and unlabeled data [Blum et al., 1998] : where the learning program is given small set of positive and negative data and a large set of unlabeled data.
2) Positive and unlabeled data [Li, et al., 2003]: where the learning program is only given the positive examples and the unlabeled Data. Our proposed methodology for text categorization is different from the above two. We can call it the third type: automatic labeled data and unlabeled data.

Bruijn [de Bruijn, 2002] provides an excellent survey on the different aspects in mining knowledge from the biomedical literature. There have mainly been four kinds of approaches used in relationship extraction from MEDLINE documents. Frequent co-occurrence approach seems to be easier and popular. Frequent co-occurrence approach focuses on the co-occurrences of two specific entity names such as disease and treatment, or protein names with a verb that indicates an association between them. Ding [Ding et al, 2002] tested co-occurrence of entities on abstracts, sentences and phrases in molecular biology articles to see which one provides the best place to identify the relations. They concluded that working with phrases gave the best precision and working with sentences gave the better recall. The second approach uses fixed regular expression linguistic templates (normally hand-crafted) [Ong et al., 2001] to search for a specific interaction verb and the surrounding entity names. Third approach uses Machine Learning techniques such as HMM [Ray et al., 2003] to learn some linguistic templates. Fourth approach [Rindflesch et al., 2000] is to try to discover relationship using a full syntactic parse and relations between syntactic components are inferred.

Our approach is a frequent co-occurrence approach but we work on co-occurrence of concepts rather than words. Our work builds on the initial work of Zhu [Zhu et al , 2003] in our group. In their work, they had presented the feasibility of using co-occurrence of MeSH terms from MEDLINE to find some useful relationships in Medical literature using ARM. Cimino's group in Columbia University has done some extensive works [Cimino et al., 1998] using co-occurrence of MeSH terms and semantic types. Their work is similar to ours and they have successfully applied this knowledge in document retrieval as well as for knowledge extraction.

Conclusion

In this paper, we have discussed some of the challenges in adaptive modeling framework for systems biology. We considered two aspects – modeling from experimental data and modeling from scientific articles text. We argued that learning of interaction and regulation maps of networks in an organism is crucial for designing an intelligent biomedical decision support system. We discussed some of the challenges involved in learning and modeling these biological networks. We explained our methodology on learning from experimental data and address why adaptive behaviour is essential for learning biological networks. Secondly, we discussed the importance of having a general adaptive system that may be able to extract knowledge

from text for diverse domain rather than one specific domain. We also discussed our ideas and efforts in building one such system.

References

1. Akutsu, T., Miyano, S. and Kuhara. S., *Identification of genetic networks from a small number of gene expression patterns under the boolean network model*. in *Pacific Sympoisum BioComputing*. 1999
2. Blaschke, C., Andrade, M.A., Ouzounis, C., Valencia, A. , *Automatic extraction of biological information from scientific text: protein-protein interactions*, Proc. Int. Conf. Intell. Syst. Mol. Biol., 30A(2) 60-67, 1999
3. Blum, A., Mitchell, T. , *Combining labeled and unlabeled data with co-training*, Proc. Of the 11th conf. on Computational Learning Theory,1998.
4. Chen, T., He, H.L. and Church, G.M.. *Modeling gene regulatory networks with differential equations.* in *Pacific Symposium on Biocomputing*. 1999. Singapore: World Scientific Publishing.
5. Cimino, J.J., Barnett, G.O., *Automatic knowledge acquisition from MEDLINE.* Methods of Information in Medical, 32(2):120-133, 1998
6. de Bruijn, B., Martin, J. International Journal of Medical Informatics, 67 1-18, 2002
7. Ding, J., Berleant,D., Nettleton, D., Wurtele, E., *Mining MEDLINE: abstracts, sentences, or phrases*, Pac. Symp. Biocomput., 326-337, 2002
8. D'haeseleer, P., et al. *Linear modeling of mRNA expression levels during CNS development and injury*. in Pacific Symposium on Biocomputing. 1999.
9. Freidman, N., et al., *Using Bayesian Networks to analyze expression data.* Journal of Computational Biology, 2000. 7: p. 601-620
10. GO database*: http://www.godatabase.org/*
11. Joshi, R., Li, X.L., Ramachandaran, S., Leong, T.Y., " *Automatic Model Structuring from Text using BioMedical Ontology*", American Association for Artificial Intelligence (AAAI) Workshop on Adaptive Text Extraction and Mining., 2004
12. KEGG database*: http://www.genome.jp/kegg/*
13. Kitano, H., *An introduction to Systems Biology.* Science, 2002. 295.
14. Li, X., Liu B. *Learning to classify text using positive and unlabeled data*, International Joint Conference on Artificial Intelligence (IJCAI), 2003.
15. Li, G. Leong, T.Y., *A framework to learn Bayesian network from changing, multiple-source biomedical data*, AAAI Spring Symposium on Challenges to decision support system in a changing world, 2005
16. MEDLINE : *http://www.ncbi.nlm.nih.gov/entrez*
17. Murphy, K. and Milan, S. , *Modeling Gene Expression Data Using Dynamic Bayesian Networks*, in *Technical Report*. 1999, MIT Artificial Intelligence Laboratory.
18. Ong, I., Glasner, J. and Page,D. , *Modelling regulatory pathways in E.coli from time series expression profiles.* Bioinformatics, 2002. 18: p. S241-S248
19. Ono, T., Hishigaki, H., Tanigami,A., Takagi,T , *Automated extraction of protein-protein interactions from the biological literature*, Bioinformatics 17(2) 155-161, 2001
20. PATHDB database*: http://www.ncgr.org/pathdb/*
21. Perrin, B.E., et al., *Gene Networks inference using dynamic Bayesian Networks.* Bioinformatics, 2003. 19: p. II138-II148
22. Rangel, C., et al., *Modeling T-cell activation using gene expression profiing and state space models.* Bioinformatics, 2004. 20: p. 1361-1372
23. Ray, S. , Craven, M., *Representing sentence structure in Hidden Markov Models for information extraction*, , International Joint Conference on Artificial Intelligence (IJCAI),, 1273-1279, 2001
24. Rindflesch, T.C., Tanabe, L., Weinstein, J.N., Hunter, L., *EDGAR: extraction of drugs, genes and relations from the biomedical literature*, IEEE Intell. Syst., 16(6): 62-67, 2000
25. SWISSProt database*: http://tw.expasy.org/sprot/*
26. Zhu, A, Li, J., Leong,T.Y., *Automated knowledge extraction for decision model construction: a data mining approach*, AMIA Annual Symposium, 2003