BIOINFORMATICS

Vol. 00 no. 00 2010
Pages 1–4

NETGEM: Network Embedded analysis of Temporal Gene Expression using Mixture models

Vinay Jethava[1], Chiranjib Bhattacharyya[1], Devdatt Dubhashi[2],Goutham N. Vemuri[3*]

[1]Computer Science and Automation Department, Indian Institute of Science, Bangalore, INDIA
[2]Department of Computer Science, Chalmers University of Technology, Göteborg, SWEDEN
[3]Systems Biology Division, Department of Chemical and Biological Engineering, Chalmers University of Technology, Göteborg, SWEDEN

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

ABSTRACT

Motivation Temporal analysis of gene expression data has been limited to identifying genes whose expression varies with time and/or correlation between genes what have similar temporal profiles. Often, the methods do not consider the underlying network constraints that connect the genes. In addition to identifying changes in the genes, it is becoming increasingly evident that interactions change substantially. Thus far, there is no systematic method to relate the temporal changes in gene expression to the dynamics of interactions between them in the context of a regulatory network. The availability of this data opens up possibilities for discovering new mechanisms of regulation and provides valuable insight into identifying time-sensitive interactions. Furthermore, such a framework would also allow for studies on the effect of a genetic perturbation on the dynamics of the interactions.

Results We present NETGEM, a tractable model rooted in markov dynamics, for analyzing temporal profiles of genetic expressions arising out of known protein interaction networks evolving with unknown dynamics. The model can do efficient inference and its parameters are learnt by a Maximum a Posteriori procedure by following a Bayesian approach. The Bayesian approach is neccesitated as the sample size of data is extremely small. When applied to real data NETGEM was successful in identifying (i) temporal interactions and determining their strength, (ii) functional categories of the actively interacting partners and (iii) dynamics of interactions in perturbed networks. One more sentence to conclude

Availability: The source code for NETGEM is available from http://www.sysbio.se/BioMet

Contact: goutham@chalmers.se

1 INTRODUCTION

Gene expression microarrays are increasingly being used to determine transcriptional regulation in response to a genetic or environmental perturbation. This has generated a vast amount of quantitative data which have compelled the use of statistical methods to identify differentially expressed genes and thereby, deduce transcriptional regulation. The general theme of these methods is to cluster genes, assuming that co-expressed genes are co-regulated. Often the inference is presented as a static network of genes that are activated or repressed by relevant transcription factors. This representation is similar to a wiring diagram of electrical circuits (Stigler *et al.*, 2007). Important parameters of regulation such as amplitude of the signal, time lag, etc in such networks can only be studied by explicitly modelling the dynamics of such a system. This has spurred interest in analyzing time series gene expression data.

Conventional methods of time series analysis maynot apply to this problem as the data has many interesting properties e.g. observations close together in time are more closely related (Glass & Kaplan., 1993). The analysis is further complicated by the fact that extremely small number of observations from different time points are available relative to variables (genes). There is the inherent risk of many genes having similar expression profile, just by random chance. Recognizing these problems, it is only recently that dedicated methods are being developed to infer temporal regulation of transcription (Leek *et al.*, 2006; Ernst & Joseph, 2006; Ramoni *et al.*, 2002), although temporal gene expression data are available much earlier. These, and other methods reviewed recently (Androulakis *et al.*, 2007) do not consider any dependency of observations between time points and hence are not suitablefor the problem at hand.

Previous methods which have focused on identifying temporal changes in the genes and/or identifying correlations in their temporal expression profiles have ignored the dynamics of interactions between them. In recent work (Song *et al.*, 2009), time-sensitive interactions were identified based on local neighbourhood selection with $L1$-regularization to obtain sparse networks. The analysis learns the topology of the network from the data, and assumes a smooth variation in the network interactions strengths to overcome the unreliability of results due to the small number of observations. This approach is extremely insightful in cases where the topology of the underlying network is unknown.

However, when network structure is known, as in the case of regulatory networks, there is an obvious benefit in incorporating this information into analyzing the dynamics of the interactions. Furthermore, it is of fundamental biological interest in determining

[*]to whom correspondence should be addressed

time-sensitive components of the networks. Currently there are no models which can be used for this purpose.

In this paper we consider the problem of learning a model from temporal profiles of genetic expressions for a regulatory network with a known structure. It is interesting to note that the dynamics has a direct bearing on the profiles but it is unobserved, and stochastic in nature. This motivates a markovian approach for building such models. However inferring a general model of the time-varying interactions turns out to be an NP-hard problem. Learning of model parameters is further complicated by the small number of observation points.

The interaction networks of baker's yeast, *Saccharomyces cerevisiae* are arguably the most well-constructed with a high level of confidence (Petranovic & Vemuri, 2009). Therefore, we used this network to study and validate our models. The genes and proteins in yeast are classified according to their biological function (Mewes *et al.*, 2007) to a high degree of resolution. This allows the possibility to relate functional classification of the network components with the temporal interactions between them.

This line of argumentation leads to two very fundamental questions: (i) can we distill observations about temporal characteristics of a group of functionally similar genes? (ii) would it be possible to model the effect of a genetic perturbation (gene deletion or addition) while comparing temporal interactions between the reference strain and its perturbed mutant?

As noted before the first question gives rise to intractable problems in a general setting. We finesse the problem of intractable inference by assuming that that interaction strengths evolve *independently* of each other. This assumption leads to a model where one can derive efficient inference procedures which have linear time complexity in number of temporal observations. To handle the problem of low sample size we advocate a Bayesian approach. Experimental results indicate that this does lead to useful models.

We extend the independent weight evolution model to solve the above mentioned problems as follows: (i) We model the evolution of interaction strength for a gene pair as a mixture of evolution characteristics of the functional categories with appropriately chosen Bayesian priors (ii) We propose a novel approach by considering that the interactions near the point of perturbation (gene deletion or addition) are affected to a greatest extent in their temporal behavior while those further away have the closer temporal profiles as in the reference strain.

This leads to the final model, Network Embedded analysis of Temporal Gene Expression data using Mixture models (NETGEM), which is used to investigate (a) inference of the time-varying interaction strengths given the limited number of observation points (b) multiple measurements for the expression levels from perturbed strains and (c) the relationship between the dynamics of interaction strengths and the functional classification of the genes.

The remainder of this manuscript is organized as follows: Section **??** describes the construction of the high confidence network. Section **??** presents our model based on the independently evolving weights assumption. We extend our model to incorporate functional categories in Section **??**. Section **??**) presents the variant for handling multiple strains. We present the experiments on synthetic and real datasets in Section **??** and conclude in Section **??**.

2 GOUTHAM'S ORIGNAL INTRO

Gene expression microarrays are used to determine transcriptional regulation, commonly in response to a genetic or environmental perturbation. They represent the snapshot of the transcription profile of all the genes in the genome. The vast amount of quantitative data generated from microarrays have compelled the use of different statistical methods to identify differentially expressed genes and thereby, deduce transcriptional regulation. The general theme of these methods is to cluster genes, assuming that co-expressed genes are co-regulated. Often the inference is presented as a static network of genes that are activated or repressed by relevant transcription factors. This representation is similar to a wiring diagram of electrical circuits [Stigler2007]. Important parameters of regulation such as amplitude of the signal, time lag, etc can only be studied using dynamic data. Analysis of time series gene expression poses additional problems since the data have a natural temporal ordering. Furthermore, analysis methodology also needs to account for the fact that observations close together in time will be more closely related [Glass1993]. Conventional methods of time series analysis cannot be borrowed for analyzing temporal gene expression data because of the small number of observations from different time points relative to variables (genes). There is the inherent risk of many genes having similar expression profile, just by random chance. Recognizing these problems, it is only recently that dedicated methods are being developed to infer temporal regulation of transcription [Leek2006, Ernst2006, Ramoni02cluster], although temporal gene expression data are available much earlier. These, and other methods reviewed recently [Androulakis2007] do not consider any dependency of observations between time points. As indicated in a recent review of the methods is available in .

The previous methods focused on identifying temporal changes in the genes and/or identifying correlations in their temporal expression profiles without considering the dynamics of the interactions between them. This has been the focus of recent work [Song2009], in which time-sensitive interactions were identified based on local neighbourhood selection with $L1$-regularization to obtain sparse networks. The analysis learns the topology of the network from the data, and assumes a smooth variation in the network interactions strengths to overcome the unreliability of results due to the small number of observations. This approach is extremely insightful in cases where the topology of the underlying network is unknown. However, when network structure is known, as in case of regulatory networks, there is an obvious benefit in incorporating this information into analyzing the dynamics of the interactions. Furthermore, it is of fundamental biological interest in determining time-sensitive components of the networks.

The objective of this paper is to develop a method that exploits knowledge of regulatory networks as well as temporal gene expression profile to determine the dynamics of the interactions. Furthermore, the method should be able to account for dependency of observations from one time point on previous time point within the constraints of the network. The method should also allow identification of the biological processes to which the time-sensitive network components belong to. Towards achieving these objectives, we use Hidden Markov Model (HMM) to capture the dependency of gene expression on time. HMMs are routinely used to analyze time course data in a wide range of applications [MacDonald1997] and more recently to analyzing gene expression data [Schliep2003, Yoneya2007]. They were used to partition time series gene expression data into clusters. In the context of our objective, we investigate a Markovian model for analyzing rewiring in a given biological network. The interaction networks of baker's yeast, *Saccharomyces cerevisiae* are arguably the most well-constructed with a high level of confidence [Petranovic2009]. Therefore, we used this organism as our model to evaluate the performance of our method. The outcome of this method is a weighted estimation of the dynamics of the interaction between components of the network. The genes and proteins in yeast are classified according to their biological function [Mewes2007] to a high degree of resolution. This allows the possibility to relate functional classification of the network components with the temporal interactions between them.

This line of argumentation leads to two very fundamental questions: (i) can we distill observations about temporal characteristics of a group of functionally similar genes? (ii) would it be possible to model the effect of a genetic perturbation (gene deletion or addition) while comparing temporal interactions between the reference strain and its perturbed mutant? We propose to use a Mixture model to address the first question. Since a gene can have multiple functionalities and hence, can belong to different functional groups simultaneously, [why are mixture models suited for this? Mention two lines introducing them]. The common approach to addressing the effect of a genetic perturbation raised in the second question is to treat the two strains separately. Here, we propose a novel approach by considering that the interactions near the point of perturbation (gene deletion or addition) are affected to a greatest extent in their temporal behavior while those further away have the closer temporal profiles as in the reference strain.

We present a Network Embedded analysis of Temporal Gene Expression data using Mixture models (NETGEM) to investigate (a) inference of the time-varying interaction strengths given the limited number of observation points (b) multiple measurements for the expression levels from perturbed strains and (c) the relationship between the dynamics of interaction strengths and the functional classification of the genes. We selected two time-series datasets of yeast gene expression in which the nutritional environment changed with time, one without any genetic perturbations and one with a deletion in a key transcription factor to demonstrate the utility of NETGEM.

3 VINAY'S ORIGINAL INTRO

Microarrays have become a routine tool in biological enquiry, geared to measure global gene expression in response to genetic or environmental perturbations. Gene expression microarrays present a snapshot of the transcriptional profile of all the genes at the time of measurement. The outcome is a vast amount of data, which has been analyzed using several statistical methods including hierarchical clustering (?), k-means clustering (?), self organizing maps (?), singular value decomposition (?). The key focus of the methods has been clustering of genes that have similar expression profile, based on the assumption that co-expressed genes are likely to be regulated. An inherent drawback of the clustering approaches is their unsuitability in the analysis of temporal expression data.

This has led to growing interest towards development of dedicated models to handle the temporal data. Deriving such models is a challenging task which is even more complicated due to the small number of observations (time points), owing to cost and/or biological limitations. Several methods have been investigated including significance analysis (?Leek *et al.*, 2006), autoregressive curves based model (Ramoni *et al.*, 2002), hidden markov models (HMM) (Schliep *et al.*, 2003; ?), mixture models (??), clustering methods (Ernst & Joseph, 2006), association rules (?). A review of the methods is available in Androulakis *et al.* (2007). However, the previous methods assume an time-invariant network topology, such as the protein-protein interaction network or the genetic network inferred from microarray data.

This paper focusses on the problem of inferring the time-varying interactions in a genetic network with known topology. In particular, We assume that the interactions network is known with a high degree of confidence [SENTENCE SAYING WHY THIS IS OK - VS BLIND LEARNING ALA SONG, BIOREF NEEDED]. The observed expression levels are controlled by a known network but the interaction strengths are varying with time. We investigate three aspects of the problem , namely, (a) inference of the time-varying interaction strengths given the limited number of observation points (b) multiple measurements for the expression levels from slightly perturbed strains, and (c) the relationship between the evolution of interaction-strengths and the functional hierarchy of the genes. [BIO SENTENCE SAYING WHY THESE ARE THE RIGHT PROBLEMS TO STUDY]. Song *et al.* (2009) studied a different version of the problem where the underlying network in unknown. In their model, the key assumption is that the interaction strength vary "smoothly" in time. We re-emphasize the two approaches are orthogonal [THIS SENTENCE NEEDS TO BE CAREFULLY PUT]

The dynamics of temporal evolution in a rewiring network is a matter of study, and is hypothesized to be stochastic in nature. In this work, we explore markovian modelling for this task. Unfortunately, in all the three cases, the associated inference problems are extremely difficult. This is further complicated by the small number of observation points. We study this problem from a Bayesian perspective, introducing suitable priors for the model parameters.

The main contribution of this paper is to develop dynamical graphical models for modeling temporal variations in interaction networks using markovian dynamics. A fully generalized treatment of the time-varying interactions leads to an NP-hard inference problem. We begin by making the simplifying assumption that the interaction strengths evolve *independently* of each other. We derive inference procedure for learning the dynamics of the interactions network. We present a theoritical justification for the assumption and present validation checks on synthetic datasets. Experiments performed on real-world datasets show promising results.

BIO-REF NEEDED We discuss a variant of the model which handles the presence of multiple strains which are slightly different perturbed versions of the original version. For example, a perturbed strain might have a couple of genes deleted compared to the base strain. Traditional methods have treated each of the slightly perturbed strains separately. However, it might be expected that in the case that the perturbed strain only slightly varies from the original strain (i.e. only a few genes are knocked out), the interactions (edges) near the knocked out genes will show a significant change in their evolution characteristics while interactions (edges) far from the knocked out genes would have the same evolution characteristics as in the reference strain.

We extend the model to incorporate the functional classifications in the analysis of temporal expression data. There has been considerable effort in establishing a hierarchy of genes based on their functionality (Bader *et al.*, 2003; Mewes *et al.*, 2002; Stark *et al.*, 2006; Xenarios *et al.*, 2000; Zanzoni, 2002). This poses the natural question of the relationship between the functional classification of the genes and the temporal evolution of the interaction. Further, can we distill some observations about evolution characteristics of a group of functionally similar genes. We model the evolution of interaction strength for a gene pair as a mixture of evolution characteristics of the functional categories.

This leads to the approximate inference algorithm, NETGEM, which models the gene interactions for a known network in terms of the functional hierarchy of the genes using the expression data over multiple strains.

3.0.1 Real world Experiments: We applied the algorithm to publicly available time-series gene expression data in Saccharomyces cerevisiae. The available of a highly curated interaction network for this organism makes it an ideal platform for testing the method. We selected two time-series datasets in which the nutritional environment changed with time, one without any genetic perturbations and one with a deletion in the $Sfp1$ transcription factor. The first dataset consists of expression of genes during the gradual transition from carbon starvation to nitrogen starvation in a D-stat under aerobic or anaerobic conditions (Farzadfard et al., 2010). Almost a fourth of the genome underwent transcriptional changes in response to the transition. The dominant transcription factor that brought about these changes was $Sfp1$, which is known to assimilate signals from the environment and coordinates growth with metabolism (Marion et al., 2004). The second dataset measures the temporal changes in gene expression upon sudden exposure of a strain of $S.cerevisiae$ in which $Sfp1$ was deleted to glucose (Cipollina et al., 2009).

The remainder of this manuscript is organized as follows: Section **??** describes the overall model, including the construction of the high confidence network (Section **??**), the factorial approximation (Section **??**), the mixture model (Section **??**) and the strain damping model (Section **??**). We present the experiments on synthetic and real datasets in Section **??** and conclude in Section **??**.

HARP ON THE EXPERIMENTS SECTION

REFERENCES

Androulakis, I. P., Yang, E. & Almon, R. R. (2007) Analysis of time-series gene expression data: methods, challenges, and opportunities. *Annual Review of Biomedical Engineering,* **9** (1), 205–228.

Bader, G. D., Betel, D. & Hogue, C. W. (2003) Bind: the biomolecular interaction network database. *Nucleic acids research,* **31** (1), 248–250.

Beal, M. J. (2003). *Variational Algorithms for Approximate Bayesian Inference*. PhD thesis, Gatsby Computational Neuroscience Unit, University College London.

Bilmes, J. (1998). A gentle tutorial of the em algorithm and its application to parameter estimation for gaussian mixture and hidden markov models. Technical report University of Washington.

Cipollina, C. *et al.* (2008) Saccharomyces cerevisiae SFP1: at the crossroads of central metabolism and ribosome biogenesis *Microbiology,* **154** , 1686–1699.

James M. Coughlan and Huiying Shen. Shape matching with belief propagation: Using dynamic quantization to accomodate occlusion and clutter. In *Generative-Model Mased Vision workshop at CVPR*, 2004.

Dempster, A. P., Laird, N. M. & Rubin, D. B. (1977) Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological),* **39** (1), 1–38.

Ernst, J. & Joseph, Z. B. (2006) Stem: a tool for the analysis of short time series gene expression data. *BMC Bioinformatics,* **7** (1).

Farzadfard, F. *et al.* (2010) Metabolic and transcriptional dynamics during the transition from carbon limitation to nitrogen limitation in saccharomyces cerevisiae. *Genome Biology (in review)*.

Felzenszwalb, P., Huttenlocher, D., and Kleinberg, J. Fast algorithms for large state space hmms with applications to web usage analysis. In *Advances in Neural Information Processing Systems*, 2003.

Friedman, N., Geiger, D., and Lotner, N. Likelihood computations using value abstraction. In *Uncertainty in Artificial Intelligence, Proceedings of the Conference on*, 2000.

Gavin, A. C., *et al.* (2006) Proteome survey reveals modularity of the yeast cell machinery. *Nature,* **440**, 631–636.

Gavin, A. C., *et al.* (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature,* **415**, 141–147.

Gelman, A. *et al.* (2003) *Bayesian Data Analysis, Second Edition (Texts in Statistical Science)*. 2 edition,, Chapman & Hall/CRC.

Gentleman, R. C. *et al.* (2004) Bioconductor: open software development for computational biology and bioinformatics. *Genome Biology,* **5** (10), R80.

Ghahramani, Z. & Jordan, M. I. (1997) Factorial hidden markov models. *Machine Learning,* **29** (2-3), 245–273.

Glass, L. & Kaplan., D. (1993) Time series analysis of complex dynamics in physiology and medicine. *Med Prog Technol,* **19**, 115–128.

Ho, Y. *et al.* (2002) Systematic identification of protein complexes in saccharomyces cerevisiae by mass spectrometry. *Nature,* **415**, 180–3.

Horn, R. A. & Johnson, C. R. (1990) *Matrix Analysis*. Cambridge University Press.

Ito, T. *et al.* (2001) A comprehensive two-hybrid analysis to explore the yeast protein interactome. *PNAS,* **98**, 4569–74.

Krogan, N. J. *et al.* (2006) Global landscape of protein complexes in the yeast saccharomyces cerevisiae. *Nature,* **440**, 637–43.

Leek, J. T. *et al.* (2006) EDGE: extraction and analysis of differential gene expression. *Bioinformatics,* **22** (4), 507–508.

MacDonald, I. L. & Zucchini, W. (1997) *Hidden Markov and other models for discrete-valued time series*. 1 edition,, Chapman & Hall, London; New York.

Mclachlan, G. J. & Krishnan, T. (1996) *The EM Algorithm and Extensions*. 1 edition,, Wiley-Interscience.

Mewes, H. W. *et al.* (2007) MIPS: analysis and annotation of genome information in 2007. *Nucleic Acids Res,* **36**.

Mewes, H. W. *et al.* (2002) Mips: a database for genomes and protein sequences. *Nucleic acids research,* **30** (1), 31–34.

Petranovic, D. & Vemuri, G. N. (2009) Impact of yeast systems biology on industrial biotechnology. *J Biotechnol,* **144** (3), 204–11.

Rabiner, L. R. (1989) A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE,* **77** (2), 257–286.

Ramoni, M. F. *et al.* (2002) Cluster analysis of gene expression dynamics. *Proc Natl Acad Sci U S A,* **99** (14), 9121–9126.

Schliep, A. *et al.* (2003) Using hidden markov models to analyze gene expression time course data. In *ISMB (Supplement of Bioinformatics)* pp. 255–263.

Shannon, P. *et al.* (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks In *Genome Research* **13** (11), pp. 2498 –2504.

Song, L., Kolar, M. & Xing, E. P. (2009) Keller: estimating time-varying interactions between genes. *Bioinformatics,* **25** (12).

Stark, C. *et al.* (2006) Biogrid: a general repository for interaction datasets. *Nucleic Acids Res,* **34** (Database issue).

Stigler, B. *et al.* (2007) Reverse engineering of dynamic networks. *Ann N Y Acad Sci,* **1115**, 168–77.

Uetz, P. *et al.* (2000) A comprehensive analysis of protein-protein interactions in saccharomyces cerevisiae. :. *Nature,* **403**, 623–7.

Xenarios, I. *et al.* (2000) Dip: the database of interacting proteins. *Nucl. Acids Res.,* **28** (1), 289–291.

Yoneya, T. & Mamitsuka, H. (2007) A hidden markov model-based approach for identifying timing differences in gene expression under different experimental factors. *Bioinformatics,* **23** (7).

Zanzoni, A. (2002) Mint: a molecular interaction database. *FEBS Letters,* **513** (1), 135–140.