BioChem. 218 – Computational Molecular Biology December 5th 2008
Molong Li

Computational Methods of Discovering Gene Modules and Constructing Regulatory Networks

Introduction

Transcriptional regulation provides cells with the critical ability to react to environmental stimuli by altering protein expression levels. It allows cells to exhibit complex behavior such as adaptation and differentiation. In eukaryotes, this is accomplished primarily by the binding of transcription factors to form protein complexes that can either activate or repress gene expression. With the ability of whole genome sequencing and high throughput data gathering, the emerging field of systems biology has begun to tackle the problem of characterizing and constructing a genome scale regulatory network. Such a network would encompass all the genes and how their products affect the transcription of all other genes. This genome scale network would allow a more comprehensive understanding of mechanisms behind cellular behavior and pave the way for whole cell modeling. However, our current knowledge of transcription factor binding sites and their effect on gene expression on a genome scale is still very limited (Abnizova et al, 2007). Computational methods of analyzing and integrating high throughput data remain the most efficient way of tackling such a large-scale problem.

It has been shown that regulatory networks are often modular in nature with a few highly connected nodes and closely interacting genes (Zhan, 2007). These gene modules are defined as sets of genes which have similar expression patterns and are bound by the same set of transcription factor (Bar-Joseph et al, 2003). Therefore, the construction of a gene regulatory network follows from the identification of gene modules and their interactions. Two recently developed high throughput technologies that have been essential to studying gene modules and regulatory networks have been microarray expression chips and chromatin immuno-precipitation on chip assays (Liu et al, 2007). The former allows for the measurement of the expression levels of all genes in the genome. More recently, the latter technology combines immuno-precipitation techniques with microarray chip technology to reveal transcription factor and DNA binding interactions.

However, the use either technology on its own has significant shortcomings. Initially, clustering algorithms were used to analyze expression data in order to find co-expressed genes.

These sets of co-expressed genes could not be described as gene modules though since one cannot definitively identify these groups of genes as being regulated by the same pathways and transcription factors (Markowetz et al, 2007). Therefore, it is possible that under conditions that have not been tested, these sets of genes would fail to co-express. On the other hand, using transcription factor binding data along says nothing of the effect on the expression of the genes. It is possible that the binding of the transcription factor does not alter expression at all. In contrast, methods of combining this heterogeneous set of data allows for the strengthening of signals in the inherently noisy and often under sampled data of microarray chips (Liu et al, 2007; Luo et al, 2007).

Thus, much work has been done to create computational algorithms for the incorporation of these two types of data, along with other genomic data, in order to accurately identify gene modules. There have been many such methods developed and one of the first such methods that moved beyond simple expression data clustering was created by Segal et al, 2003. Their method only incorporated expression data with motif data. Since then, more complex algorithms have been developed and this study will present five of these methods to gain insight into their scope and evolution (Table 1).

Name	Description	Reference
GRAM	Searches for co-bound genes with a strict cutoff. Then relaxes cutoff for genes that co-express with the original set.	Bar-Joseph et al, 2003
SAMBA	Discretizes expression and binding data into gene properties. Algorithm then looks for genes with statistically significant common property sets.	Tanay et al, 2003
ReMoDiscovery	Stringent and relaxed two step procedure that combined motif, expression, and ChIP-chip data.	Lemmens et al, 2006
COGRIM	Uses a Bayesian network to model expression level as a function of transcription factor expression and binding.	Chen et al, 2007
Inferelator	Uses biclustering to group co-expressed genes and then machine learning to infer regulatory influence from RNA and protein expression levels.	Bonneau et al, 2006

Table 1.
A list of relevant computational methods for combining heterogeneous data sets for gene module identification and regulatory network construction.

Methods

GRAM (Genetic Regulatory Modules) Algorithm: Bar-Joseph et al, 2003

The GRAM algorithm was developed to integrate microarray expression data and ChIP-chip binding data as Bar-Joseph et al. viewed these as complementary pieces of data. In doing so, they hoped to reduce the level of noise inherent in microarray data and accurately identify gene modules of co-expressed and co-bound genes. The algorithm beings by exhaustively identifying all combinations of transcription binding and found the sets of genes that corresponded to these common transcription factor bindings. The initial criteria for determine a binding event was set to be stringent, (p-value threshold of .0001) in order to reduce the incidence of false positives at the expense of false negatives. A second step incorporating the expression data using relaxed transcription factor binding thresholds that would recover the false negatives without increasing false positives. This process involves determining the "center" of the expression data for the sets of genes in each tentative module. This could be done by a variety of metrics such as Euclidean or angle distance where the dimensions of the space corresponded to the number of conditions in the expression data. Using this expression center, one could remove genes from the module which were too dissimilar in expression and add genes that were co-expressed with a lower transcription binding threshold (p-value threshold of .001). Thus, one would have modified the original transcription binding module resulting in a final module whose genes were co-expressed and co-bound with false positive rates better than if one had only used one set of data alone.

This method allows genes and transcription factors to belong in more than one module and connecting the interactions between modules could lead to the construction of regulatory networks. Although this type of analysis cannot interpret dynamic data, it was able to identify possible complex regulatory mechanism such as feed forward loops. These were characterized by a transcriptional regulator binding to a gene that produced another regulator, both of which would then bind a set of common genes. However, the inability of this method to take advantage of time sequence expression data was a notable shortcoming (Ernst et al, 2007). It is often the case that two gene who are being regulated by the same transcriptional factor may exhibit opposite transcriptional influence (either activating or repressing). The statistical and static approach by which this method incorporates expression data prevents it from predicting activation or inhibiting relationships between transcription factors and regulated genes (Li et al, 2008). In addition, the p-value thresholds are rather arbitrary (Chen et al, 2007). However, more recent methods have used more sophisticated algorithms to overcome these shortcomings and have shown to outperform the GRAM algorithm in identifying functionally relevant gene modules (Liu et al, 2007).

SAMBA (Statistical-Algorithmic Method for Bicluster Analysis): Tanay et al, 2003

SAMBA takes an entirely different approach to incorporating heterogeneous data sets. First, it converts all information about the genes such as expression data and binding data into generic gene properties. They then represent the set of genes and gene products as a weighted bipartite graph in which a gene is linked to a gene property via a probability score which corresponds to the relevance of the gene property to that particular gene. The consequence of this uniform, discretized representation of genomic data is that it allows for a wide variety of data sets besides expression and binding data. The study included data such as protein interactions and growth phenotypes. A biclustering algorithm is then applied to the bipartite graph to identify statistically significant subgraphs of gene modules. Biclustering is a method by which the rows and columns of a matrix may be clustered simultaneously. In this application, it allows for the identification of subgroups of genes with similar gene properties across a subset of the gene properties. This method also allows overlapping modules which is useful for network reconstruction.

The primary advantage of this algorithm, the ability to incorporate a wide range of data types, is also a shortcoming. In the process of homogenizing data types into gene properties, it discretizes data that may be inherently continuous, such as gene expression. There is no generalized method for this discretization, leading to the loss of data content (Liu et al, 2007). In addition, there are more recent biclustering algorithms that have been developed which are more robust and can handle multiple data sets (Reiss et al, 2006).

ReMoDiscovery: Lemmens et al, 2006

The study describes this method as unique in that it incorporates heterogeneous data sets concurrently (non-iteratively) into the model. In addition, it uses motif data as an independent data source rather than as a method of checking results. The method starts with a seed discovery step in which stringent criteria are used to find gene modules. This includes a high degree of co-expression, a minimum number of common transcriptional regulators, as well as a minimum number of conserved motifs. Rather than exhaustively searching for gene sets, the algorithm only looks for “maximal modules” which are defined as seed modules which become invalid with the addition of any other gene. This greatly reduces the computational costs of the algorithm. Once these seed modules have been identified, there is a seed extension step. This involves ranking the remaining genes by their similarity in expression profile to the mean expression profile of the seed set. Cutoffs for this ranked list are determined by iteratively choosing cutoffs and measuring the module enrichments of all motifs. The optimal cutoff is one with maximal enrichment values.

One benefit of this method is its ability to rank genes within each module. This provides a metric for the likelihood that the inclusion of gene in a module is biologically relevant. However, it seems that this method also relies heavily on motif data, both for seed discovery and extension. Since there are other methods of determining cutoffs during the seed extension step, such as optimizing for functional enrichment of the genes in the set through functional ontology, it would be interesting to see if this different parameter would affect the identification of gene modules. Overall, a notable shortcoming of this method is the arbitrary nature of the cutoffs (Chen et al, 2007).

COGRIM (Clustering Of Genes into Regulons using Integrated Modeling): Chen et al, 2007

Chen argues that the previously mentioned algorithms all use arbitrary threshold parameters and require knowledge about the contributions of each various data sets. To overcome these problems, the COGRIM method was developed which uses Bayesian networks to integrate and model the different data sets. Bayesian networks are probabilistic models in which probabilistic distributions of output are a function of a set of variables. In this case, the framework of the Bayesian network allows for the linear regression of gene expression data as a function of binding data. By entering in the heterogeneous data sets: transcription binding data, ChIP binding data, and sequence level binding data, the constructed regulatory model is able to probabilistically predict four sets of genes for every transcription factor: gene targets predicted by the model and binding data alone, gene targets predicted by the model but not binding data alone, gene targets predicted by binding data alone but no by the model, and finally, genes not predicted as targets by either the model or binding data alone. This model framework also allows estimations of interactions between transcription factors, either synergistic or antagonistic.

In Chen's study, he included a performance comparison between COGRIM, GRAM, and ReMoDiscovery. The study used all three methods to identify gene modules in a yeast model and compared the within module expression correlation as well as functional enrichment. The study found that the COGRIM method performed better than both previous models due to its probabilistic nature whereas previous methods used arbitrary threshold parameters. Though Chen's study used proteins as regulatory factors, the Bayesian network approach does not preclude using other elements as regulatory factors. One shortcoming of Bayesian networks is that they are in the form of directed acyclic graphs and therefore cannot represent feedback control mechanisms. Compared to other network representations, such as deterministic differential equations and stochastic modeling, Bayesian networks lose a lot of the network complexity. However, this is made up for by robustness and ease of interpretation (Bonneau et al, 2006).

Inferelator: Bonneau et al. 2006

The Inferelator is an additive linear model which integrates expression data with genome annotation to infer regulatory influences. It uses a form of supervised machine learning in which it is given a training set in order to learn the regulatory network. It can then be used to predict gene expression in novel scenarios. Compared to stochastic modeling and Bayesian networks, this method is intermediate as far as complexity and robustness. Additive linear models predict gene expression levels using a weighted sum of its function predictors. Decreasing the dimensionally of data sets greatly reduces the task of a learning algorithm. Therefore, a biclustering algorithm, cMoneky (Reiss 2006), is used to group co-regulated genes into clusters before using the Inferelator. By providing the algorithm with a training set of expression data, it was able to infer gene regulatory influences from RNA and protein expression. This model is able to model both steady state and time course expression levels, as well as learn causal relationships through gene knockouts and time course simulations.

While this method is very promising, there are still significant steps needed to perfect the method. For one, the algorithm's modeled regulatory networks can only be interpreted as suggested regulatory interactions that may not resolve indirect mechanisms of regulation. Unlike some of the previous methods for identifying gene modules, this method does not incorporate direct experimental evidence of transcription factor and gene pair-wise interactions from ChIP-

chip data. Rather than using direct measurements of transcription factor proteins levels, the method infers its models from RNA expression levels which add additional levels of error. The ability of this method to predict expression levels have so far been limited to testing using cross validation, an iterative process in which subsets of the data is used as training sets and validation sets. By incorporating accurate global data sets as well as ChIP-chip data, this method has a lot of potential. Whereas the methods described previously are limited to identification of gene modules, the Inferelator has taken the next step: regulatory network construction and predictions of gene expression.

Discussion

Much work has been down over the past few years in the study of regulatory network reconstruction via integration of heterogeneous high throughput data. It is certainly a difficult endeavor and these five examples provide an overview of the various paths taken to tackle the problem of gene module identification and network construction. While each of these methods has it pros and cons, the COGRIM method and Inferelator both have a lot of potential and are among the latest developments in the field. The chief advantage of the COGRIM Bayesian network approach is its abandonment of arbitrary threshold parameter, thereby giving it a performance edge over previous method of gene module identification. The Inferelator on the hand uses machine learning and additive linear modeling to produce expression level predictions that have been very accurate. Though it still requires modification, it has brought the field closer to accurately modeling cell behavior via a constructed regulatory network. Other examples of machine learning techniques have already begun emerging to improve upon the Inferelator method.

Applications for regulatory network reconstruction are wide ranging. An accurate genome scale regulatory network could provide insight into new drug targets as well as the ability to model drug effectiveness. In addition, it may be possible to create synthetic networks for industrial biotechnology purposes (Hayete et al, 2007). It is also one step closer to whole cell modeling, an achievement that would truly be astounding.

References

Abnizova I, Subhankulova T, Gilks W. Recent computational approaches to understand gene regulation: mining gene regulation in silico. Curr Genomics. 2007 Apr;8(2):79-91.

Bar-Joseph Z, Gerber GK, Lee TI, Rinaldi NJ, Yoo JY, Robert F, Gordon DB, Fraenkel E, Jaakkola TS, Young RA, Gifford DK. Computational discovery of gene modules and regulatory networks. Nat Biotechnol. 2003 Nov;21(11):1337-42. Epub 2003 Oct 12.

Bonneau R, Reiss DJ, Shannon P, Facciotti M, Hood L, Baliga NS, Thorsson V. The Inferelator: an algorithm for learning parsimonious regulatory networks from systems-biology data sets de novo. Genome Biol. 2006;7(5):R36. Epub 2006 May 10.

Chen G, Jensen ST, Stoeckert CJ Jr. Clustering of genes into regulons using integrated modeling-COGRIM. Genome Biol. 2007;8(1):R4.

Elkon R, Linhart C, Sharan R, Shamir R, Shiloh Y. Genome-wide in silico identification of transcriptional regulators controlling the cell cycle in human cells. Genome Res. 2003 May;13(5):773-80.

Ernst J, Vainas O, Harbison CT, Simon I, Bar-Joseph Z. Reconstructing dynamic regulatory maps. Mol Syst Biol. 2007;3:74. Epub 2007 Jan 16.

Faith JJ, Hayete B, Thaden JT, Mogno I, Wierzbowski J, Cottarel G, Kasif S, Collins JJ, Gardner TS. Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles. PLoS Biol. 2007 Jan;5(1):e8.

Hayete B, Gardner TS, Collins JJ. Size matters: network inference tackles the genome scale. Mol Syst Biol. 2007;3:77. Epub 2007 Feb 13.

Lemmens K, Dhollander T, De Bie T, Monsieurs P, Engelen K, Smets B, Winderickx J, De Moor B, Marchal K. Inferring transcriptional modules from ChIP-chip, motif and microarray data. Genome Biol. 2006;7(5):R37. Epub 2006 May 5.

Li H, Zhan M. Unraveling transcriptional regulatory programs by integrative analysis of microarray and transcription factor binding data. Bioinformatics. 2008 Sep 1;24(17):1874-80. Epub 2008 Jun 27.

Liu X, Jessen WJ, Sivaganesan S, Aronow BJ, Medvedovic M. Bayesian hierarchical model for transcriptional module discovery by jointly modeling gene expression and ChIP-chip data. BMC Bioinformatics. 2007 Aug 3;8:283.

Luo F, Yang Y, Zhong J, Gao H, Khan L, Thompson DK, Zhou J. Constructing gene co-expression networks and predicting functions of unknown genes by random matrix theory. BMC Bioinformatics. 2007 Aug 14;8:299.

Markowetz F, Spang R. Inferring cellular networks--a review. BMC Bioinformatics. 2007 Sep 27;8 Suppl 6:S5.

Price ND, Shmulevich I.Biochemical and statistical network models for systems biology. Curr Opin Biotechnol. 2007 Aug;18(4):365-70. Epub 2007 Aug 3.

Reiss DJ, Baliga NS, Bonneau R. Integrated biclustering of heterogeneous genome-wide datasets for the inference of global regulatory networks. BMC Bioinformatics. 2006 Jun 2;7:280.

Segal E, Shapira M, Regev A, Pe'er D, Botstein D, Koller D, Friedman N. Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. Nat Genet. 2003 Jun;34(2):166-76.

Tanay A, Sharan R, Kupiec M, Shamir R. Revealing modularity and organization in the yeast molecular network by integrated analysis of highly heterogeneous genomewide data. Proc Natl Acad Sci U S A. 2004 Mar 2;101(9):2981-6. Epub 2004 Feb 18.

Zhan M. Deciphering modular and dynamic behaviors of transcriptional networks. Genomic Med. 2007;1(1-2):19-28. Epub 2007 May 11.