Chapter 1

Remembrance of Experiments Past: A redescription based tool for discovery in complex systems

Samantha Kleinberg
NYU Bioinformatics Group, Courant Institute
Marco Antoniotti
NYU Bioinformatics Group, Courant Institute
DISCo, Università Milano-Bicocca, Italy
Satish Tadepalli
Computer Science, Virginia Tech
Naren Ramakrishnan
Computer Science, Virginia Tech
Bud Mishra
NYU Bioinformatics Group, Courant Institute

A complex system creates a "whole that is larger than the sum of its parts," by coordinating many interacting simpler component processes. Yet, each of these processes is difficult to decipher as their visible signatures are only seen in a syntactic background, devoid of the context. Examples of such visible datasets are time-course description of gene- expression abundance levels, neural spike-trains, or click-streams for web pages. It has now become rather effortless to collect voluminous datasets of this nature; but how can we make sense of them and draw significant conclusions?

For instance, in the case of time-course gene-expression datasets, rather than following small sets of known genes, can we develop a holistic approach that provides a view of the entire system as it evolves through time?

We have developed GOALIE (Gene-Ontology for Algorithmic Logic and Invariant Extraction) - a systems biology application that presents global and dynamic perspectives (e.g., invariants) inferred collectively over a gene-expression dataset. Such perspectives are important in order to obtain a process-level understanding of the underlying cellular machinery; especially how cells react, respond, and recover from environmental changes. GOALIE uncovers formal temporal logic models of biological processes by redescribing time course microarray data into the vocabulary of biological processes and then piecing these redescriptions together into a Kripke structure. In such a model, possible worlds encode transcriptional states and are connected to future possible worlds by state transitions. An HKM (Hidden Kripke Model) constructed in this manner then supports various query, inference, and comparative assessment tasks, besides providing descriptive process-level summaries. The formal basis for GOALIE is a multi-attribute information bottleneck (IB) formulation, where we aim to retain the most relevant information about states and their transitions while at the same time compressing the number of syntactic signatures used for representing the data. We describe the mathematical formulation, software implementation, and a case study of the yeast (*S. cerevisiae*) cell cycle.

1.1 The problem of microarray analysis

Microarrays, which allow the measurement of expression levels for tens of thousands of genes at a time, are a useful technique for gathering biological data, but it can be difficult to make sense of the results they produce. Experiments can be repeated with varying conditions (such as starvation, heat shock, etc.) and with each microarray having upwards of 10,000 probes, and many time course experiments having over 10 time points, there is a vast amount of data being generated. One frequent method used to deal with this is to cluster[2] the data into groups that have similar properties. There are two common ways of doing this clustering: by expression patterns over the entire dataset, which fails to take into account variations of the data over time; and by function - using a known ontology of biological processes, which fails to find unknown groupings. What is needed is a method of modeling the data based on both biological function and temporal evolution and a way to relate it to other experiments.

We want to make inferences about the system, simplifying it for ease, while taking care to retain its most important parts. But how can we determine what is important in the system, without predetermining our results? In many cases, due to the size and complexity of microarray output, it is only possible to follow small sets of known genes and then work outward, attempting to identify more. Our approach draws the biologists' attention to specific sets of genes and processes that appear to be interesting based on mathematical computations - not their own prior knowledge, while allowing a framework for exploring these and other items in further detail as they relate to the whole dataset. By making certain abstractions, while leaving the essential underlying structure of the experiment

intact, we provide the best of both worlds.

1.2 Computation

Our computational methods are based on a temporal redescription, which takes genes and translates them into a controlled vocabulary, and then stitches those translations together to form a picture of the biological system as it evolves over time. To facilitate the construction of our model of the dataset, we begin by breaking the data into small overlapping windows of time. Each window contains all of the genes in the data set, but with only their expression values for a specific interval of time. The initial cluster analysis, using the numerical gene expression data, is done on these windows, rather than the entire dataset. By doing this "windowing," we have simplified each computational step while also allowing for the fact that groups of genes may briefly act together but diverge across the whole dataset. Using this windowed approach we can make inferences such as "process A and process B act together beginning in hours 1-2 and continuing through hours 2-4. They are then joined by process C during hours 4-6." Each cluster in each window of time is first redescribed into the vocabulary of biological processes, and then we redescribe the clusters again: in relation to each other. That is, we connect clusters across time windows by tracking their describing terms. The connection relationships can be both one-to-many and many-to-one, as their meaning is that the terms in the connection persist from the source cluster to the destination cluster. A more rigorous algorithm for constructing these clusters and cluster-connections has been developed using a generalization of Shannon-Kolmogorov's rate-distortion theory, called "information bottleneck approach:" in this setting, the redescription is simply viewed as a lossy-compression of all the temporal observations by a simpler finite automaton that introduces minimal distortion in terms of the known ontologies. This generalized algorithm appears elsewhere[4].

The basis for the computations we perform is the use of temporal logic in the form of a Kripke structure. Here, this is a directed graph (often acyclic, DAG), defined by its vertices, edges, and properties (V, E, P). The vertices represent the reachable states of the system (clusters), edges are transitions between states (cluster connections across time windows), and properties (GO Categories) are used to annotate the states in which they are true. Terms within the Gene Ontology (GO)[6] have their own hierarchical structure, which is also incorporated into the model. In the case of our yeast (*S. cerevisiae*) data, describing terms would include "cell cycle," and more specifically, "M phase" and "G1/S transition of mitotic cell cycle."

1.2.1 Computation in detail: HKM and IB

This section explains in detail, the methods used in GOALIE to derive a Kripke model in the form of a DAG from a given time series data set.

The Information Bottleneck Principle

The Information Bottleneck principle[12] is an information theoretic approach to clustering. Suppose each instance x_i of a random variable X is associated with an instance y_i of another random variable Y, and we desire a clustered representation T of X which preserves as much information about the relevant variable Y as possible. Typically X represents the features of the objects to be compressed and Y represents some relevant information about their classes. According to the IB principle the clustering scheme T is chosen to minimize the functional

$$\mathcal{L}_{min} = I(T;X) - \beta I(T;Y) \tag{1.1}$$

where $I(T;X)$ represents the mutual information between T and X, $I(T;X) = \sum_{t,x} \log \frac{p(t,x)}{p(t)p(x)}$, and β is the Lagrange parameter that controls the tradeoff between compression and preservation of relevant information. Lower values of β give more importance to compression while higher values give more importance to preserving relevant information. We can try to maximize the functional $\mathcal{L}$ in equation 1.1 as follows

$$\mathcal{L}_{max} = I(T;Y) - \beta^{-1} I(T;X) \tag{1.2}$$

This equation provides a metric to evaluate different clustering schemes of X.

The key steps in GOALIE are

1. Partitioning the given time series data into windows.
2. Clustering expression data in each window.
3. Connecting the clusters in neighboring windows

Clustering within a window

Let G represent the random variable corresponding to the expression vectors in a given window W. Each expression vector g_i is an instantiation of G. Still *et al.* [11] show that k-means clustering algorithm can be derived from the information bottleneck method. Following equation 1.2, the IB formulation of k-means clustering C is as follows

$$\mathcal{L}_{max} = I(C;G) - \beta^{-1} I(C;i) \tag{1.3}$$

The clustering C compresses the data indices i while preserving similarity in expression space G.

Choosing the partitioning of windows

There are many ways to define an objective criterion for achieving a partitioning of the windows. One approach is as follows. For a given time series data set, we identify a tiling (overlapping or non-overlapping) of windows such that when each

window is re-stated in terms of clusters, the cluster identities in one window are informative of cluster identities in the neighboring window. The sum of mutual information across all windows along with a penalty for overlaps can then be used to define the IB functional at this stage.

Connecting the clusters in neighboring windows

Given the set of windows that span across all the time points in the data set, our next task is to connect the clusters in neighboring windows to track the temporal relationships in the data. In each window we find the GO ids enriched using the Fisher-Exact test with Benjamini-Hochberg (or with an empirical Bayesian approach) correction. Two clusters C_i and C_j are said to be $\theta - equivalent$ if the Jaccard's coefficient between the sets of GO ids enriched in C_i and C_j is $\geq \theta$. With these connections between the clusters, we can introduce a directed graph $G = (C, E)$ whose vertices V are the clusters and an edge from cluster C_i to C_j exists if they are in neighboring windows and the Jaccard's coefficient of the GO ids in the clusters is at least θ.

1.3 Software and Results

The main output of the software is a representation of the HKM as a directed graph. The graph presents clusters and their connections and allows exploration of the genes and terms comprising each. Integrated links to websites such as Entrez[5] and the Affymetrix database allow further study of the probes. We present a validation process for GOALIE that was tested primarily using the yeast (S. cerevisiae) cell cycle dataset collected by Spellman et al. [9]. There are three components to the Spellman yeast cell cycle data, following the yeast's behavior with α-factor, cdc15 and elutriation. This paper describes a case study using the alpha factor dataset. Further discussion of the visualization and software facilities in GOALIE is the subject of a forthcoming paper.[1]

1.3.1 Cluster Graph

We began with time course gene expression data for 6178 genes at 18 time points. After filtering out genes with missing expression data, we ended up with 4489 genes left. The resulting data was partitioned into 5 time windows, with all but the last window having an overlap of two time points. Each window was divided into 15 clusters, yielding a total of 75 clusters. These clusters formed the initial input to GOALIE. They were re-described using the Fisher-Exact test with Benjamini-Hochberg correction to control false discovery rate and a p-value of 0.05. The redescription across time windows was computed with a Jacquard's coefficient $\theta = 0.8$. Using the notation W:C to denote cluster number C in window W, we describe a sampling of our results, as seen in figure 1.1.

1:4 to 2:15 and 2:3 DNA replication initiation as well as DNA replication checkpoint are initially up regulated, consistent with the S phase of the cell cycle.

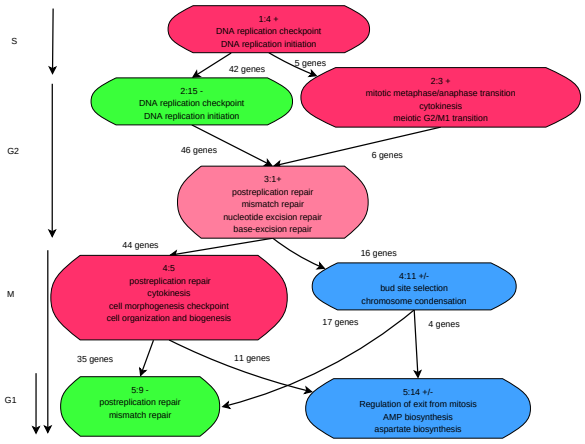

Figure 1.1: Diagram of GOALIE's output of the yeast cell cycle

1:4 then splits into two clusters, with opposite regulation patterns. In cluster 2:15, notice that the processes labeling the S portion of 1:4 are down regulated, while cytokinesis, meiotic G2/M1 transition and mitotic metaphase/anaphase transition are up regulated in 2:3.

3:1 At this time, genes from 2:15 and 2:3 which had been regulated separately converge, as G2 continues and processes associated with repairs are up regulated. As in the first window, this cluster again splits into two groups.

4:5 Cell organization and biogenesis, cell morphogenesis checkpoint, cytokinesis and postreplication repair are all up-regulated, signaling the beginning of the M phase. Genes from his cluster divide into oppositely regulated groups in the last window

4:11 Also part of M phase, but more sharply regulated, Bud site selection and chromosome condensation are first high and then go down in window 4. This cluster then separates into two groups, which overlap with genes from 4:5.

5:9 and 5:15 As The M phase ends and G1 begins, 5:9 contains down regulated processes such as post replication repair and mismatch repair. In 5:14, there is a positive regulation of the exit from mitosis, AMP biosynthesis and aspartate biosynthesis as the cell readies for the G1 phase.

1.3.2 Gantt chart view

Another way of viewing the data is through the use of a Gantt chart[3], which is a bar graph designed for the display of data with a time component. The use of Gantt charts simplifies the creation of summaries of datasets, allowing users to view either all terms describing clusters or just a few selected terms. The chart displays the state of the regulation of biological terms across the time windows. The average expression level (i.e. up, down, normal, inactive) of a term and its components (i.e., descendants in the GO hierarchy) is used to determine the

corresponding color codes (Red, Green, Yellow, and Black) in the display. This information is computed via the cluster centroids for each of the terms in the windows in which it appears. By looking at the chart, one can obtain an overall sense of the time course evolution underlying the dataset and infer patterns in the activity of groups of genes. There is an inherent information loss in this method, but it still provides a complement to the main cluster graph.

1.4 Comparison of datasets

Often, we wish not just to examine the patterns within a microarray dataset, but also consider the more interesting question of how they may relate to another dataset. For example, one may compare a cancer cell line against the same cell line in the presence of a growth or some other extra-cellular factor. In this case, it can be difficult to determine the pertinent similarities and differences between experiments just from a cursory examination of the clustering results. We are exploring a possible solution based on aligning the Gantt charts of the experiments. In this approach, immediately after creating the Gantt charts for each experiment and converting the expression profiles of the terms to strings (where up, down, inactivity, and normal activity within a time window are each represented by a character), the terms are aligned to one another using a dynamic programming pairwise alignment algorithm with a similarity-scoring matrix. The resulting alignment emphasizes matches, regardless of gaps, and strongly penalizes anti-correlated expression (i.e., situation when in one dataset term A is always up, while in another dataset it is always down). By presenting the results ordered by score, it is simple to visually filter out terms that are consistent across datasets, while highlighting those that diverge. Additionally, the terms are sorted such that they form groups based on similar expression patterns (similar within each dataset, not across the datasets). In this way, users are immediately presented with possible terms of interest and insight into patterns of biological processes.

1.5 Conclusion and future directions

Many complex systems, whether natural or engineered, are amenable to GOALIE's semantic analysis within the kinds of logical frameworks it creates. Through the examples used here, such a logical framework has been seen to provide a new way of reasoning about complex biological systems as well as the interface that makes this information accessible to scientists. In the future, GOALIE is planned to provide support for other ontologies and controlled vocabularies, such as MeSH [8] and KEGG[7]. There are also several interesting technical questions to be answered: How does one select the optimal size of the Kripke model, mostly determined by window size and number of clusters[10], as those choices can strongly affect the rate (data-compression), distortion, and hence, the fidelity of the resulting models. Additionally, GOALIE will need to

continually evolve to interface with the users from other fields through transparent representations such as the Gantt charts, which minimize the information loss and provide more background information on the genetic basis for the displayed terms.

1.6 Acknowledgements

We would like to acknowledge the funding support for GOALIE from the DARPA BioCOMP program, Dept. of Homeland Security, and the NSF EMT program.

Bibliography

[1] Antoniotti, Marco, Samantha Kleinberg, Satish Tadepalli, Naren Ramakrishnan, and Bud Mishra, "On ontology-based enrichment, summarization and visualization of time-course gene expression data", Under review (2006).

[2] Bar-Joseph, Ziv, "Analyzing time series gene expression data", *Bioinformatics* **20**, 16 (2004), 2493–2503.

[3] Clark, Wallace, *The Gantt Chart* third ed., Pitman and Sons (1952).

[4] Cover, Thomas M., and Joy A. Thomas, *Elements of information theory*, John Wiley and Sons (1991).

[5] "Entrez pubmed", `http://www.ncbi.nlm.nih.gov/entrez/query.fcgi`.

[6] "Gene ontology consortium site", `http://www.geneontology.org`.

[7] "Kegg database", `http://www.genome.ad.jp/kegg/`.

[8] "MeSH site", `http://www.nlm.nih.gov/mesh/meshhome.html` (1999).

[9] Spellman, Paul T., Gavin Sherlock, Michael Q. Zhang, Vishwanath R. Iyer, Kirk Anders, Michael B. Eisen, Patrick O. Brown, David Botstein, and Bruce Futcher, "Comprehensive Identification of Cell Cycle-Regulated Genes of the Yeast Saccharomyces Cerevisiae by Microarray Hybridization", *Molecular Biology of the Cell* **9** (1998), 3273–3297.

[10] Still, Susanne, and William Bialek, "How Many Clusters? An Information-Theoretic Perspective", *Neural Computation* **16** (2004), 2483–2506.

[11] Still, Susanne, William Bialek, and Léon Bottou, "Geometric clustering using the information bottleneck method.", *NIPS*, (2003).

[12] Tishby, Naftali, Fernando C. Pereira, and William Bialek, "The Information Bottleneck Method", *Proceedings of the 37th Annual Allerton Conference on Communication, Control and Computing*, (1999).