Logic in the Time of Malaria: Segmenting Time Course Data to Understand the *Plasmodium Falciparum* Life Cycle

Samantha Kleinberg, Kevin Casey, & Bud Mishra
Courant Institute of Mathematical Sciences, New York University

Abstract

GOALIE (GO Algorithmic Logic for Information Extraction) tools were developed for high throughput biological data, such as time-course gene expression microarray data, and to enable a detailed look at the activities of biological systems and their temporal evolution. The underlying mathematical questions are challenging, as the biological processes being studied may comprise many sub-processes, each with their own underlying rules. Processes may be initiated and then briefly co-regulated before diverging; they may also take place across varying time scales. Finding the relationships between components of the system, as well as characterizing the core features of the process taking place, can be difficult tasks given the volume of data and the subtlety of interactions.

Our approach to the problem combines tools from information theory, model checking, and logic. To this end, we have developed and implemented many new solutions for these problems within GOALIE, an automated tool for reconstructing temporal models of biological systems. Using a variant on the Information Bottleneck (IB) method we find critical time-points, which define an optimal segmentation of the data into time windows. These critical time-points are the location of significant process level reorganization. Then, we may form testable hypotheses about the system using the built in logical structure. In this work, we applied GOALIE to the study of *Plasmodium Falciparum's* Intraerythrocytic Developmental Cycle(IDC).

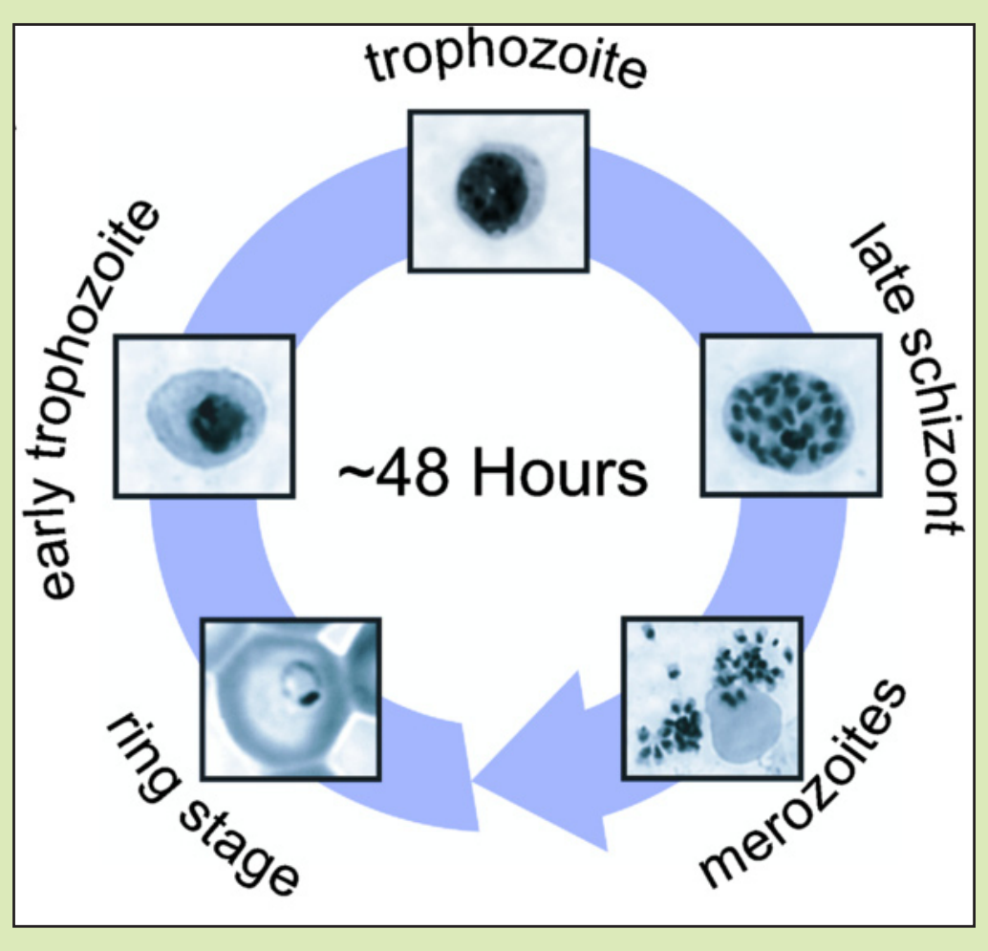
Plasmodium Falciparum

What is it?

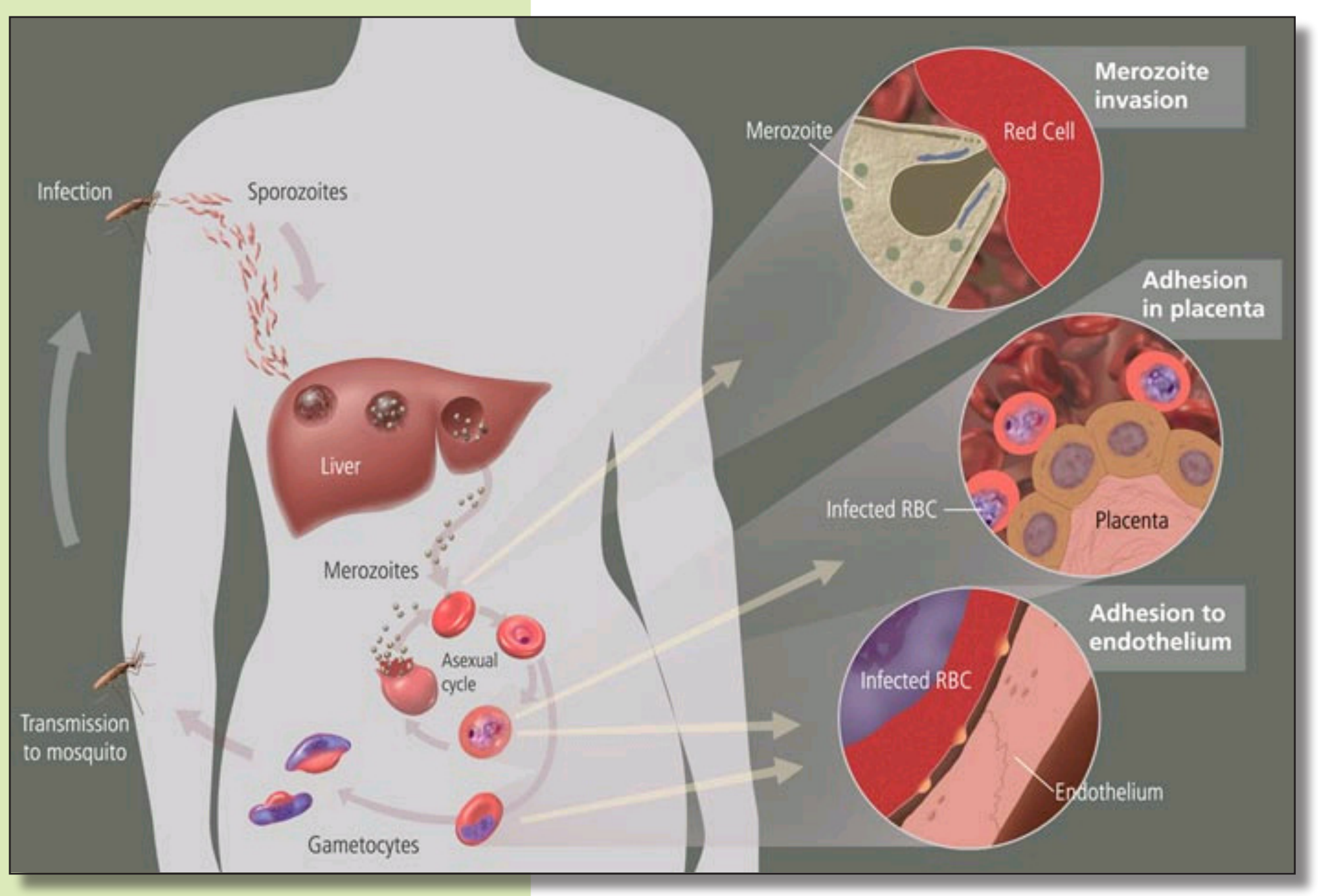
- A strain of *Plasmodium* that causes a deadly form of chloroquinine-resistant malaria
- Results in up to 2 million deaths/year
- No current vaccine, or long term solution.

Why study it?

- Its genome was recently sequenced.
- There are 5,400 genes, with few known functions.



IDC of *Plasmodium Falciparum*.
Bozdech et al. The Transcriptome of the Intraerythrocytic Developmental Cycle of Plasmodium falciparum. PLoS Biology 1/1/2003 e5



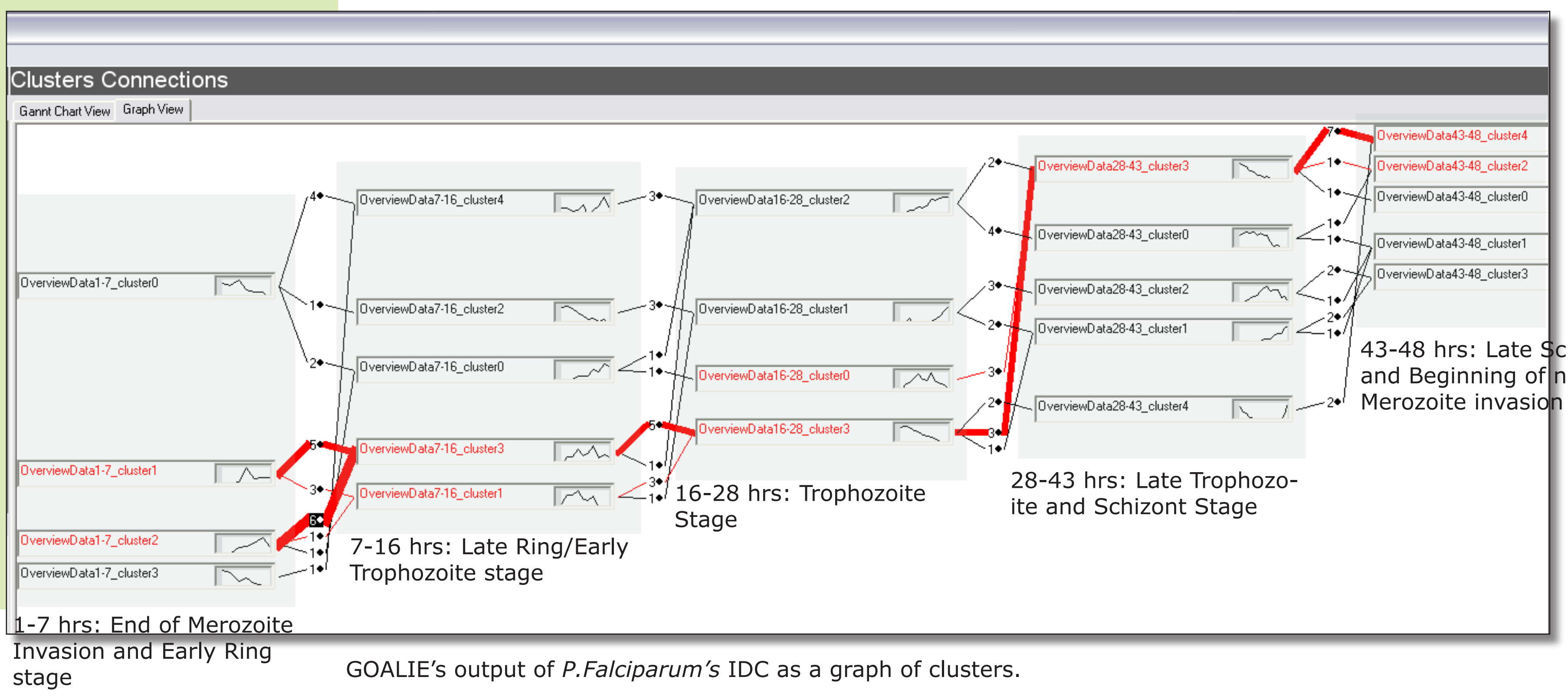
NIH diagram of the Malaria parasite life-cycle

Intraerythrocytic Developmental Cycle(IDC)

- Blood stage has three substages, forming the IDC: Ring, Trophozoite, and Schizont stages
- Much of the *P. Falciparum* genome is active during the IDC – as one set of genes is deactivated, another is being turned on
- Contains stages of life cycle responsible for malaria symptoms

Results

- We analyzed microarray data covering 48 hours with 43 data points
- Each microarray consisted of 3719 oligonucleotides (2714 Open Reading Frames), of which 1878 were annotated with GO terms
- Even with ~50% annotation and no knowledge of the underlying structure, we were able to find the important features of the system
- The windows found compared well with the known main IDC stages
    - Critical time points: 7, 16 28 and 43 hours.
    - 17 hrs and 29 hrs are known transitions from ring to trophozoite and trophozoite to schizont respectively.
- Clustering showed that most genes/processes involved in the three substages were co-regulated across the entire time course
- Recovered the cascade of activity - clusters correspond to groups of genes being up- and then down-regulated in turn


Clusters Connections
Gannt Chart View
Graph View
OverviewData1-7_cluster0
OverviewData1-7_cluster1
OverviewData1-7_cluster2
OverviewData1-7_cluster3
OverviewData7-16_cluster4
OverviewData7-16_cluster2
OverviewData7-16_cluster0
OverviewData7-16_cluster3
OverviewData7-16_cluster1
OverviewData16-28_cluster2
OverviewData16-28_cluster1
OverviewData16-28_cluster0
OverviewData16-28_cluster3
OverviewData28-43_cluster3
OverviewData28-43_cluster0
OverviewData28-43_cluster2
OverviewData28-43_cluster1
OverviewData28-43_cluster4
OverviewData43-48_cluster4
OverviewData43-48_cluster2
OverviewData43-48_cluster0
OverviewData43-48_cluster1
OverviewData43-48_cluster3
1-7 hrs: End of Merozoite Invasion and Early Ring stage
7-16 hrs: Late Ring/Early Trophozoite stage
16-28 hrs: Trophozoite Stage
28-43 hrs: Late Trophozoite and Schizont Stage
43-48 hrs: Late Schizont and Beginning of next Merozoite invasion

GOALIE's output of *P.Falciparum's* IDC as a graph of clusters.

GO terms related to the Ring stage have been selected. Their path through the clusters is shown in red. Note that they are co-regulated through the entire time course.

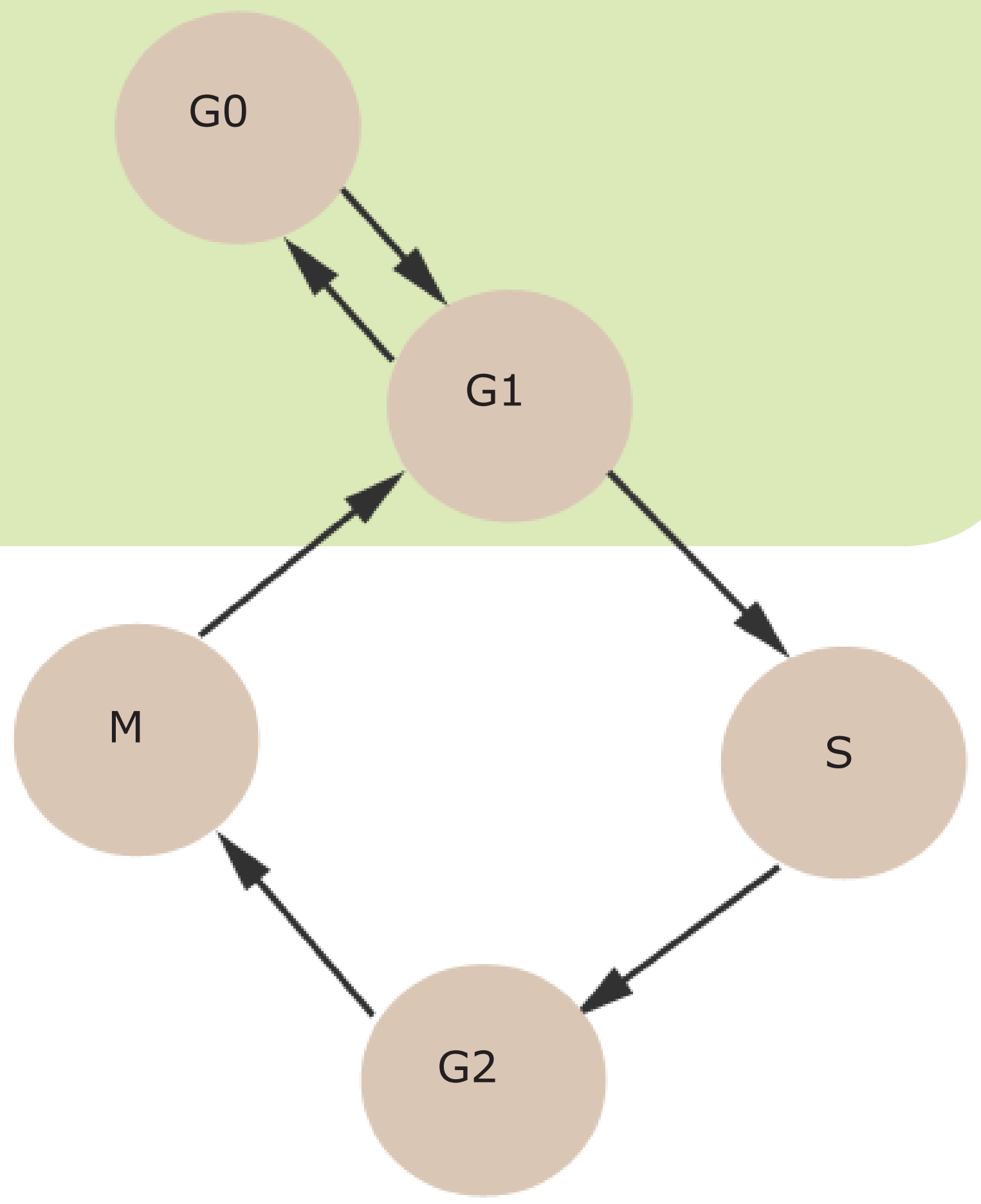
Methods

Information theory

- Genes may be briefly co-regulated, then diverge, but clustering the entire time course at once does not capture this behavior
- Define optimal windows as the ones that allow maximal (lossy) compression of data (as constrained by a distortion term).
- Find critical time points by reformulating tradeoff between compression and information as a graph search problem and use shortest path algorithm
- Use these critical time points to break data into windows containing data for all genes in the data set for a subset of all time points
- Cluster within windows using mutual information

Temporal Redescription

- Translate genes into controlled vocabulary, i.e. Gene Ontology(GO)
- Track biological processes as they move across windows, using them to connect clusters
- Summarize connections with Kripke structure
- Kripke structure is comprised of:
  Verties: reachable states of the system
  Edges: Transitions between states
  Properties: Labels denoting properties true within the states
- Here, states are clusters, edges connect clusters and labels are GO terms
- Can use built in logic of Kripke structure to ask questions about pathways through system
- Example: Starting when G1 is true, is it possible to reach M without going through G2?



Kripke structure representation of the Yeast cell cycle

References

Bar-Joseph, Z.: Analyzing time series gene expression data. Bioinformatics 20(16)(2004) 2493–2503

Bozdech, Z., Llinas, M., Pulliam, B., Wong, E., Zhu, J., DeRisi, J.: The transcriptome of the intraerythrocytic developmental cycle of. Plasmodium falciparum. PLoS Biol 1 (2003)

Casagrande, A., Casey, K., Falch, R. Piazza1, C., Rupert, B., Vizzotto, G., Mishra, B.: Translating time-course gene expression profiles into semi-algebraic hybrid automata via dimensionality reduction? Submitted (December 2007)

Clarke, E.M., Grunberg, O., Peled, D.A.: Model Checking. MIT Press (1999)

Kleinberg, S., Antoniotti, M., Tadepalli, S., Ramakrishnan, N., Mishra, B.: Remembrance of experiments past: a redescription approach for knowledge discovery in complex systems (2006)

Slonim, N., Atwal, G.S.S., Tkacik, G., Bialek, W.: Information-based clustering. Proc Natl Acad Sci U S A (December 2005)

Further Information

Please email mishra@nyu.edu or visit http://bioinformatics.nyu.edu/Projects/GOALIE/ for more information on GOALIE