Using Scan Statistics for Anomaly Detection in Genetic Regulatory Networks

Christopher C. Overall
Department of Bioinformatics & Computational Biology
College of Science
George Mason University

QMDNS 2007
2/8/2007

Collaborators: J.L. Solka, Carey E. Priebe, J.W. Weller

Introduction

- Biological systems contain many levels of complex interactions between heterogeneous components (DNA, RNA, protein), the dynamics of which are usually non-linear.
- Often modeled in two ways:
  - Dynamical model using systems of differential equations
  - Static network (graph) model
- The dynamical models can be computationally intensive or intractable due to the large number of differential equations (hundreds or thousands).

Introduction (Cont'd)

- It has become increasing popular to represent biological interactions as a network (graph) in which a node (vertex) represents a biological molecule or functional complex and an edge represents a relationship between the two molecules. Simple representation, harness power of graph theory, but lose dynamics.
- Biological researchers often want to determine if and when the relationship between biological entities alters significantly over time i.e. anomaly detection.
- Analogous to anomaly-based network intrusion detection in the computer network security domain.

Introduction (Cont'd)

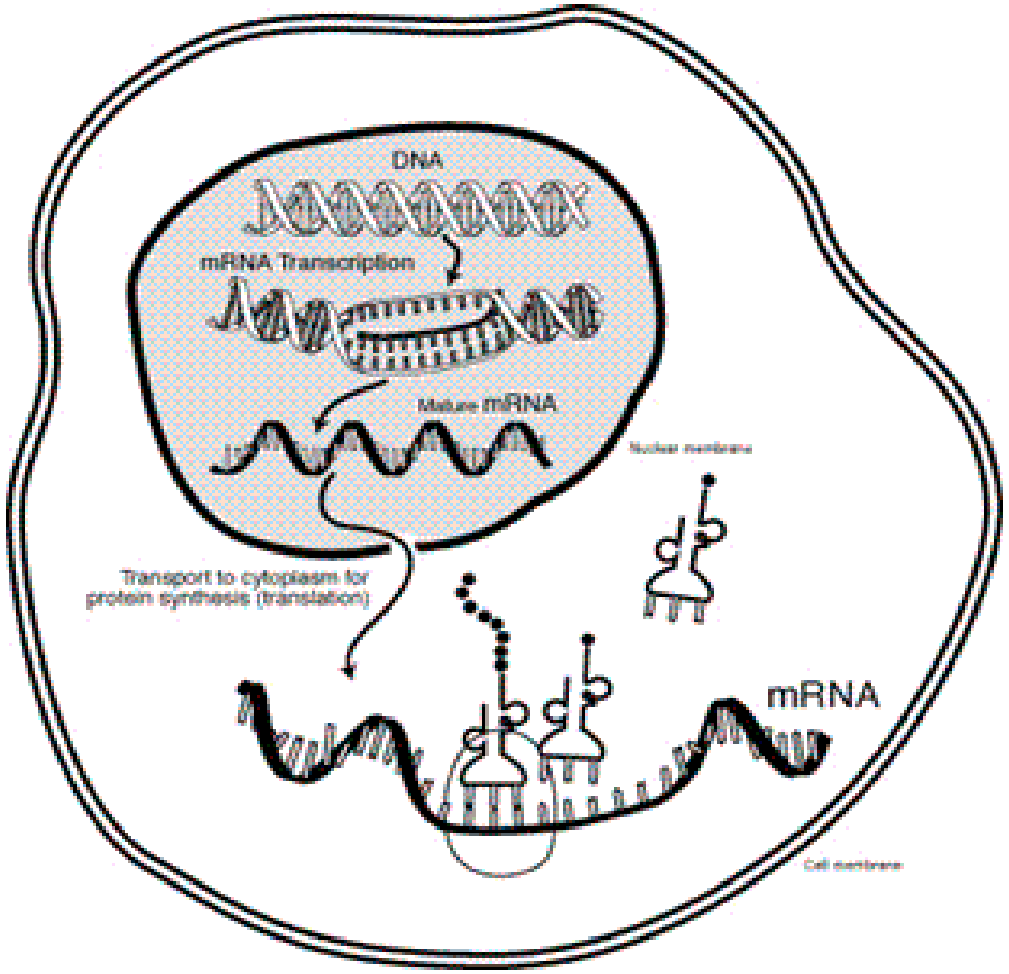
- There is a large amount of data that requires automated techniques for determining normal network behavior and then using this prior history to determine when an anomalous change has occurred at one or more nodes in the network (when and where).
- These techniques detect anomalous behavior in the network that might not have been deduced by a human, allowing the analyst or researcher to hone in on the anomaly and to determine its significance.
- To this end, in the biological domain, we need anomaly detection techniques that use simplifying representations of the biological interactions but also maintain some of the dynamics.

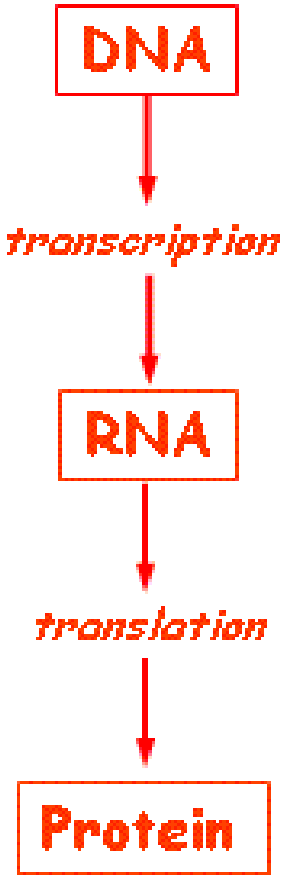
Introduction (Cont'd)

- Our anomaly detection methodology is a hybrid solution (simplifying representation and incorporation of some dynamics) that determines when a gene changes its level of co-expression with other genes over time (anomalous behavior), which may signal an important cellular event centered about that gene.

A Very Brief Introduction to Relevant Biological Concepts

Central Dogma of Molecular Biology

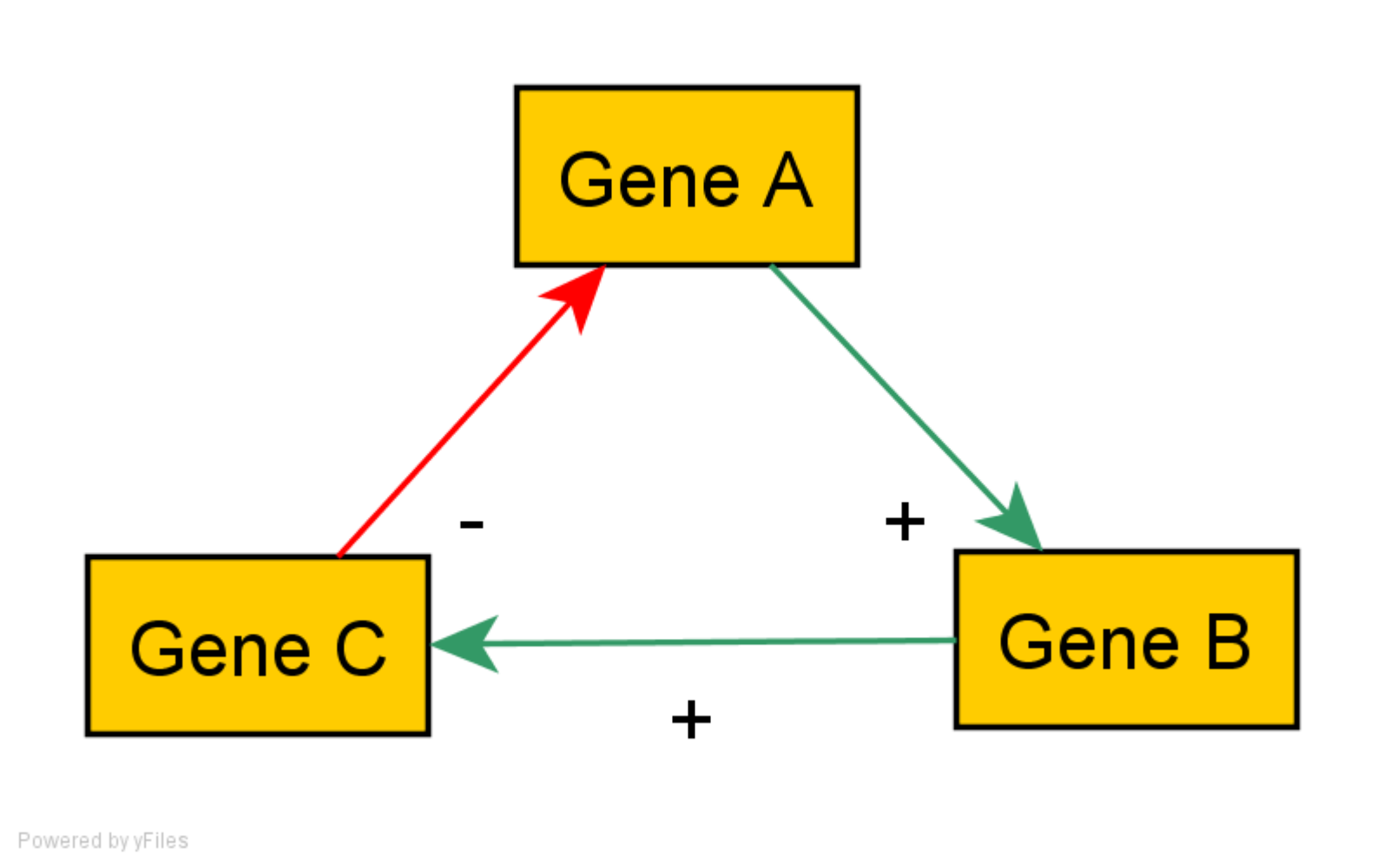




http://www.ncbi.nlm.nih.gov/Class/MLACourse/Modules/MolBioReview/central_dogma.html

A Very Brief Introduction to Relevant Biological Concepts

What is a Genetic Regulatory Network?

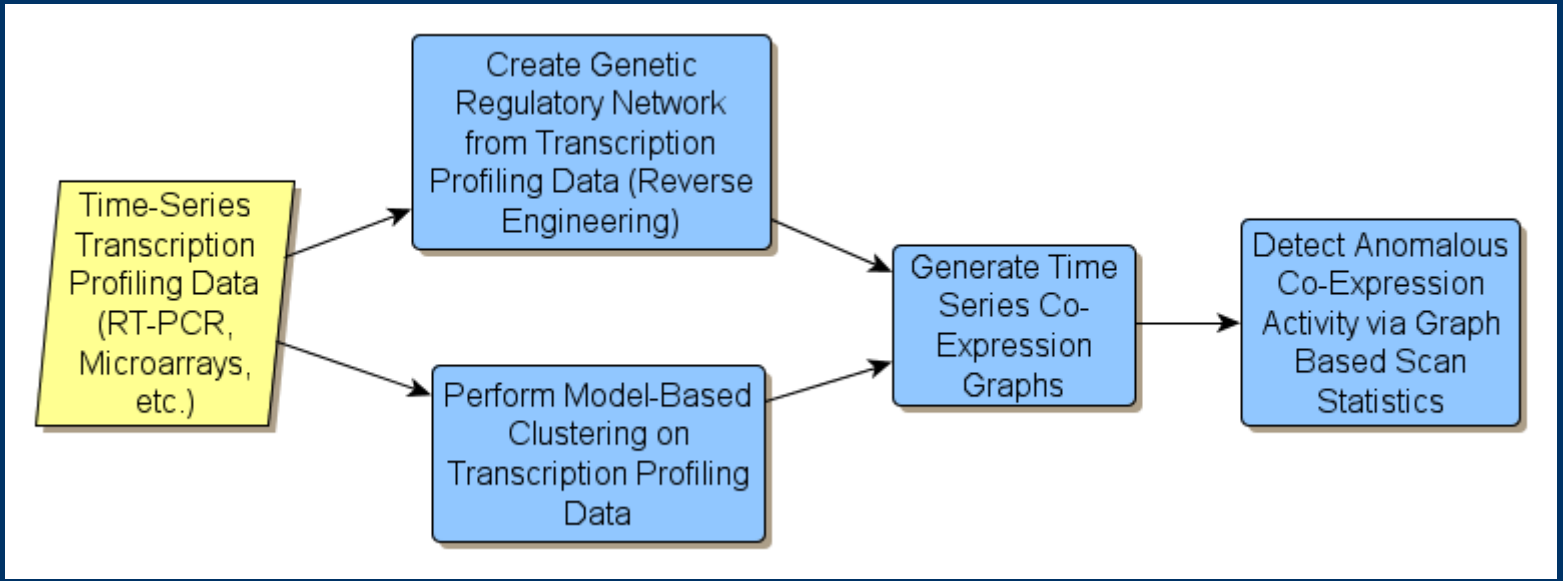


A Very Brief Introduction to Relevant Biological Concepts

Transcription Profiling

- Methodologies that are used to measure the abundance of transcripts (mRNA) in a cell under certain conditions e.g. microarrays and RT-PCR.
- Often a comparison of a test sample vs. a reference sample e.g. cancerous cells vs. normal cells or a comparison over multiple time points during a particular process? Is the test case induced/suppressed/no change compared to the control?
- Provides insight into which genes or combination of genes are expressed differently in the different conditions which may provide molecular targets for new medicines or diagnostic tools.

Approach

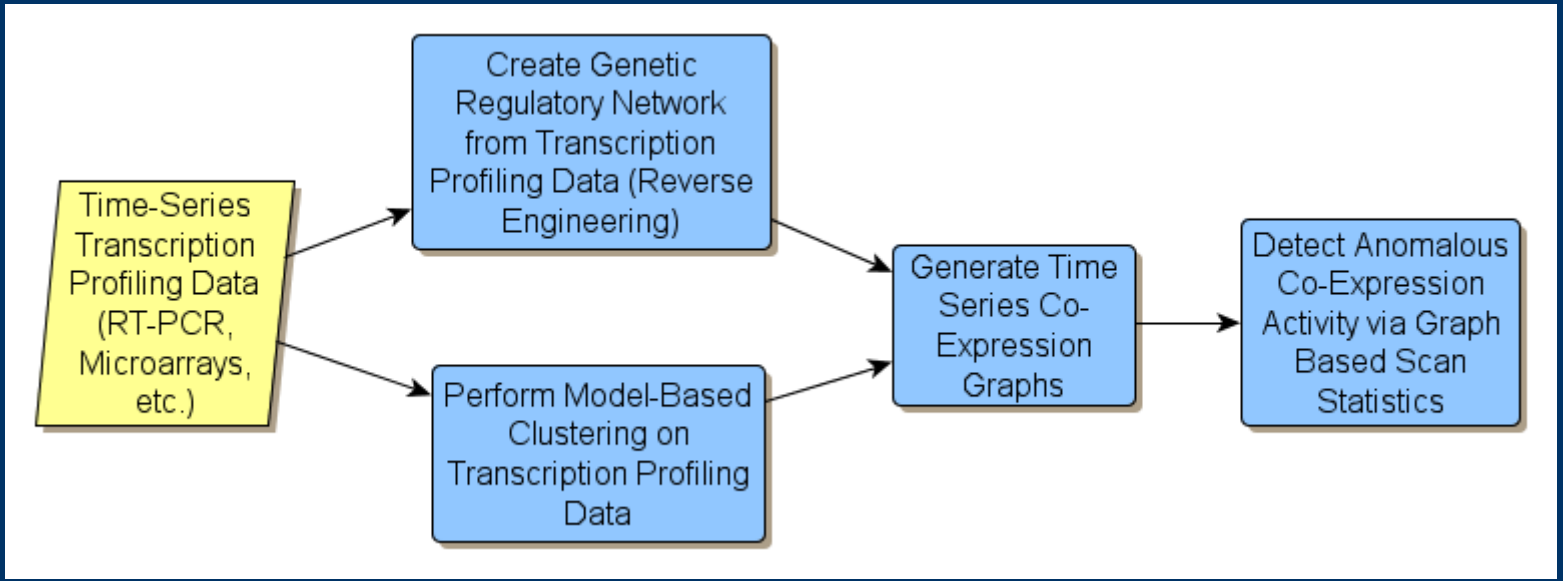

Time-Series Transcription Profiling Data (RT-PCR, Microarrays, etc.)
Create Genetic Regulatory Network from Transcription Profiling Data (Reverse Engineering)
Perform Model-Based Clustering on Transcription Profiling Data
Generate Time Series Co-Expression Graphs
Detect Anomalous Co-Expression Activity via Graph Based Scan Statistics

Time-Series Transcription Profiling Dataset

- **Drosophila Development**: cDNA microarray Drosophila development dataset from Arbeitman et al. (2002). There were 66 total time points, spanning 4 developmental stages: embryonic, larval, pupal, and adult. We considered a subset of the total number of genes (132) during the 18 time points gathered in the pupal stage.

Approach

Model of Genetic Regulatory Network

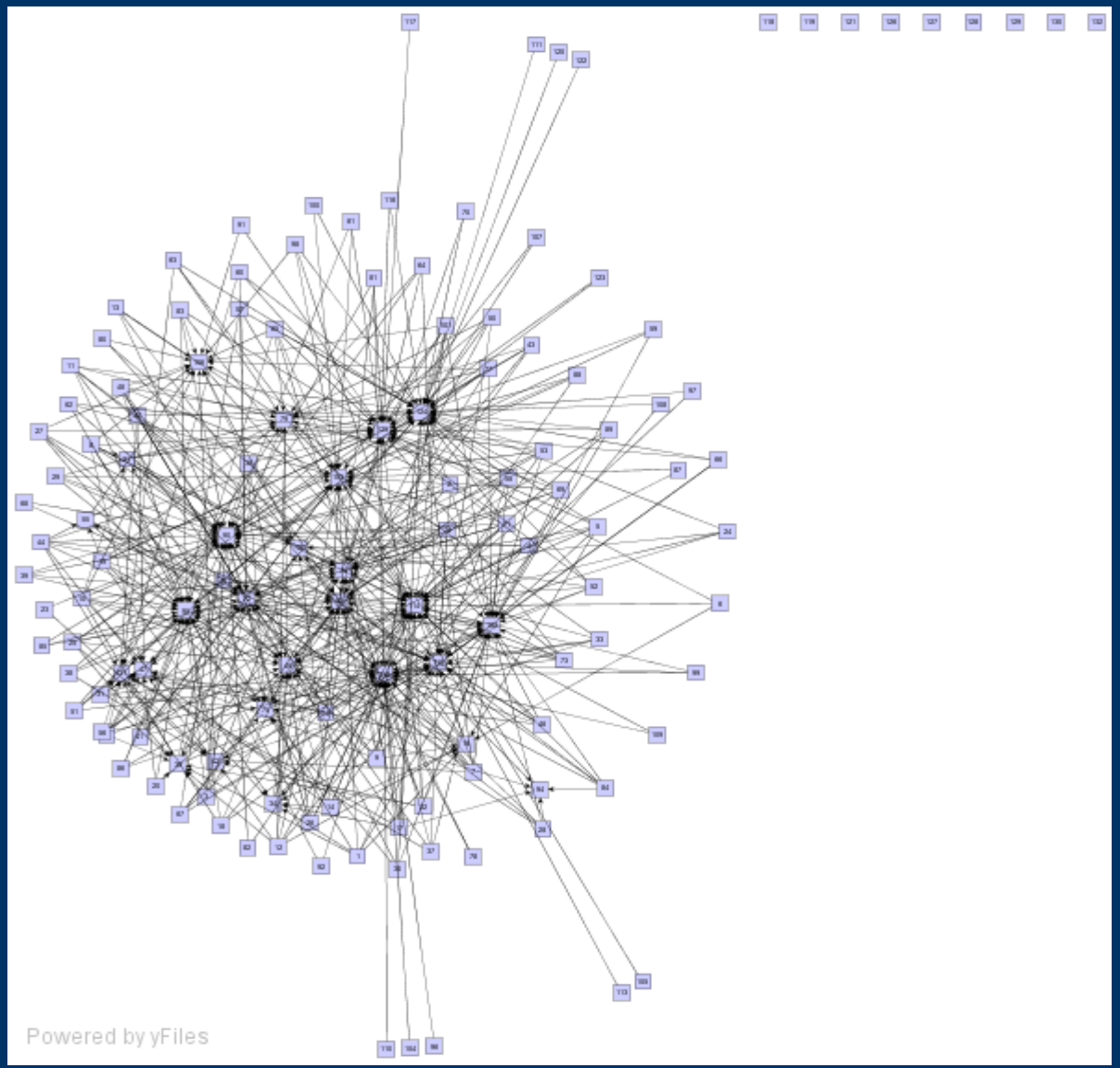


Model of Genetic Regulatory Networks from Transcription Profiling Data

- Reverse engineering methodologies.
- Different inference models, including:
  - Linear Model: uses linear differential equations to model the interactions
  - S-system: rate law approximations with non-linear ordinary differential equations
  - Boolean Network: the relationships between the genes are represented as ON/OFF.
  - Dynamic Bayesian Network

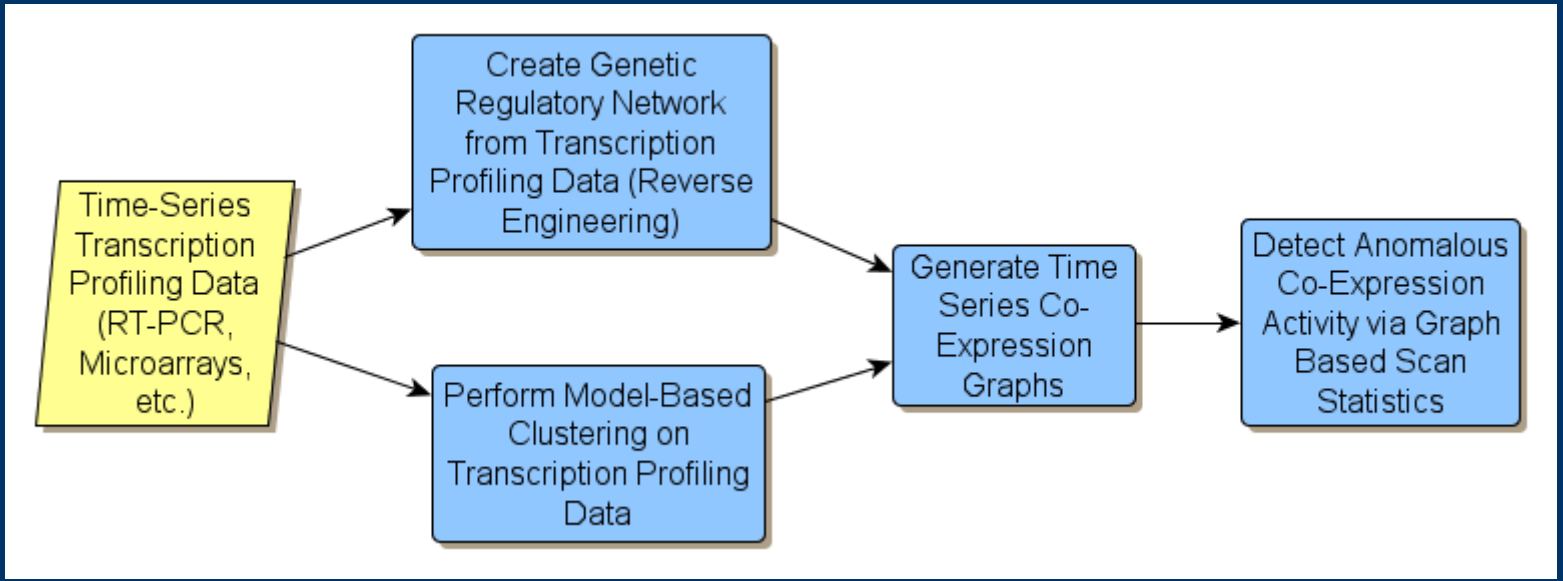
Model of Genetic Regulatory Network - Example

Arbeitman Dataset



Approach

Perform Model-Based Clustering on Dataset



Model-Based Clustering

- Clustering is based on a probability model.
- The data (independent samples) are assumed to be generated from a mixture of probability distributions e.g. Gaussian.
- Univariate and multivariate.
- Clustering is reduced to a model selection problem.
- Bayesian Information Criterion (BIC) is often used for model selection.
- The posterior probabilities for membership in each cluster are generated for the data points and used to assign a data point to its final cluster.

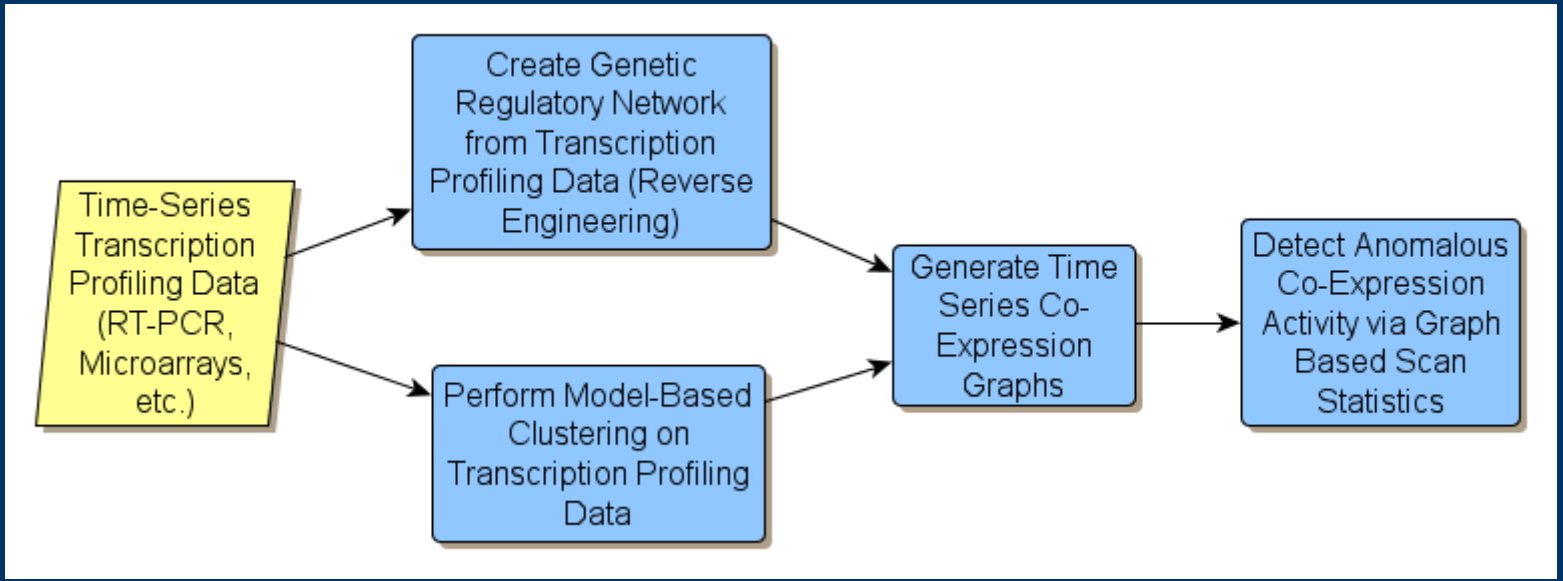
Model-Based Clustering on Transcription Profiling Data

- Group the genes at each time point by performing **multivariate model-based clustering** on the expression values over a specified time range.
- Two genes are **putatively co-expressed** if one of the following is true:
  - They are in the same cluster (similar expression pattern)
  - They are in different clusters, but one of the following is true:
    - abs(P[g1 in C1] – P[g2 in C1]) ≤ φ
    - abs(P[g1 in C2] – P[g2 in C2]) ≤ φ

    where φ = probability difference threshold.
- If the two genes are **putatively co-expressed**, we move on to the next step. Otherwise, move on to the next gene pair.

Approach

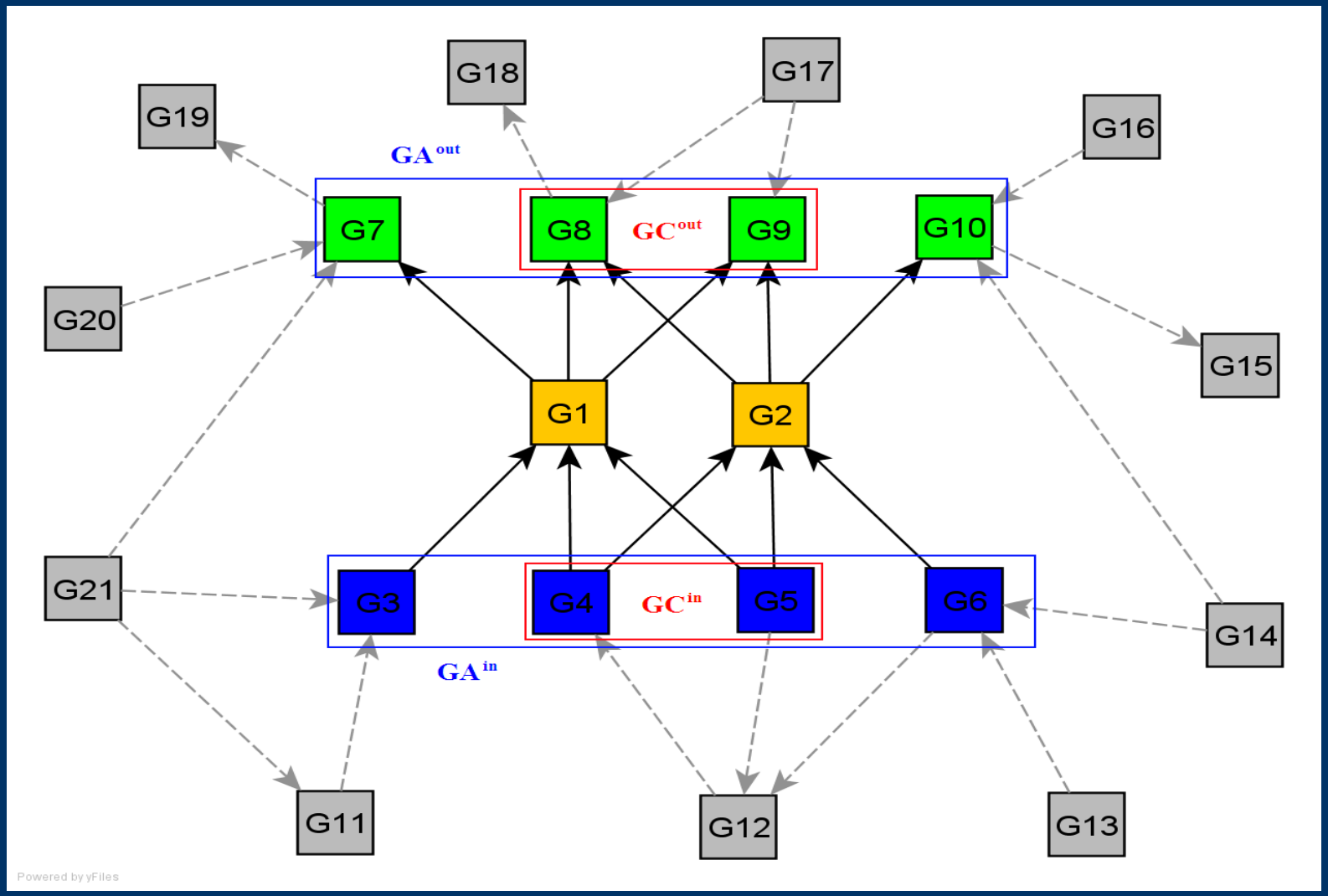
Generate Time-Series Co-Expression Graphs



Generate Time-Series Co-Expression Graphs

- Now we will use the model of the genetic regulatory network to determine if the putatively co-expressed genes can be classified as co-expressed for that time point.
- Regulators = the set of genes that regulate the expression level of a particular gene
- Regulatees = the set of genes whose expression levels are regulated by a particular gene

Generate Time-Series Co-Expression Graphs



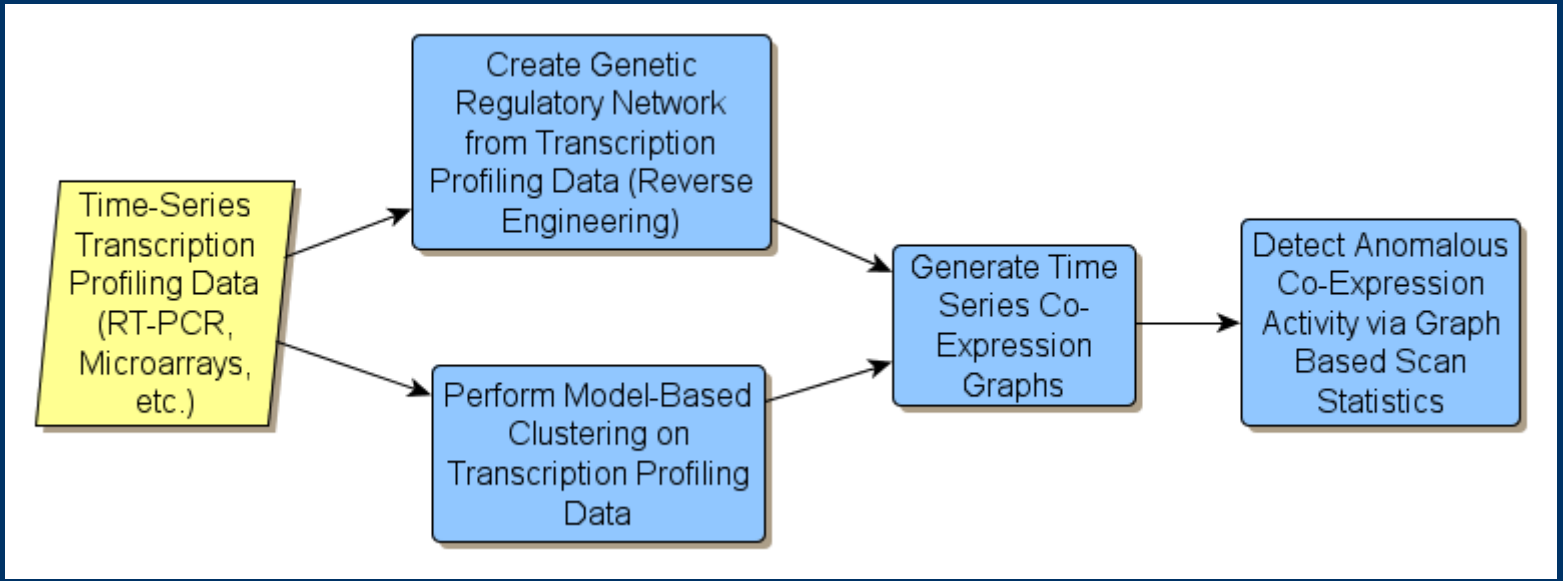
Generate Time-Series Co-Expression Graphs

- Create a regulation similarity score for each gene pair in a time point as follows:

$$Similarity = \frac{|GC_{\text{in}}| + |GC_{out}|}{|GA_{\text{in}}| + |GA_{out}|}$$

- This will give a similarity score between zero and one.
- If the similarity score is greater than or equal to a similarity threshold, λ, then we consider them to be co-expressed and draw an edge between them.
- This is repeated for every gene pair in a particular time point. Looping through all time points in the dataset, we get a co-expression graph for each time point.

Approach


Time-Series Transcription Profiling Data (RT-PCR, Microarrays, etc.)
Create Genetic Regulatory Network from Transcription Profiling Data (Reverse Engineering)
Perform Model-Based Clustering on Transcription Profiling Data
Generate Time Series Co-Expression Graphs
Detect Anomalous Co-Expression Activity via Graph Based Scan Statistics

Scan Statistics

- Commonly used to investigate an instantiation of a random field X (e.g. image of pixel values) for the presence of a local signal.
- "Moving window analysis" - scan a small window over the data and calculate a local statistic (e.g. average pixel value) for each window.
- Scan Statistic - maximum of the locality statistics, denoted M(x).
- Under a specified "homogeneity" null hypothesis H_0 on X (e.g. Gausian random field), specify a critical value c_α such that $P_{H0}[M(X) \geq c_\alpha] = \alpha$.
- If M(x) >= c_α, one can infer that there exists a nonhomogeneity – a local region with statistically significiant signal.

Graph-Based Scan Statistics

Introduction

- Priebe et al. (2005) extended the use of scan statistics to directed graphs.
- Succesfully detected anomalous activity in a time-series of social networks derived from the publicly available Enron email dataset.
- The detections ranged from email aliasing to balanced “chatter” in the second-order neighborhood about a central person.

Graph-Based Scan Statistics

Methodology

- Generate a graph for each time point in the dataset.
- Loop through each time point in the dataset.
- For each time point graph, do the following:
  - Count the number of edges in the k-neighborhood of each vertex v to get the "scale-specific" locality statistics, $\tilde{\Psi}_{k,t}(v)$
  - Normalize the locality statistics over a time period tau, to get the vertex-standardized locality statistics, $\tilde{\Psi}_{k,t}(v)$
  - Find the maximum $\tilde{\Psi}_{k,t}(v)$ to get the standardized scan statistic for each graph, $\tilde{M}_{k,t}$
- Normalize $\tilde{M}_{k,t}$ over a time period L to get the temporally-normalized standardized scan statistic, $S_{k,t}$
- $S_{k,t} \geq 5$ indicates a detection (nonhomogeneity)

Graph-Based Scan Statistics

Indicator Functions

- The **non-zero baseline** indicator function is given as

$$\tilde{\Psi}_{\mathrm{k,t}}(v) \cdot I\{\hat{\mu}_{0,t,T}(v) > c\}$$

- The **balanced "chatter"** indicator function is given as

$$\tilde{\Psi}'_{\mathrm{k,t}}(v) = \tilde{\Psi}_{2,t}(v) \cdot A_{t,\tau}(v) / max(\gamma_t(v), 1)$$

where A is the product of three indicator functions, as follows:

$$I\{\hat{\mu}_{0,t,\tau} > c_1\}$$

$$I\{\Psi_0(v) < \hat{\sigma}_{0,t,\tau}(v)c_2 + \hat{\mu}_{0,t,\tau}(v)\}$$

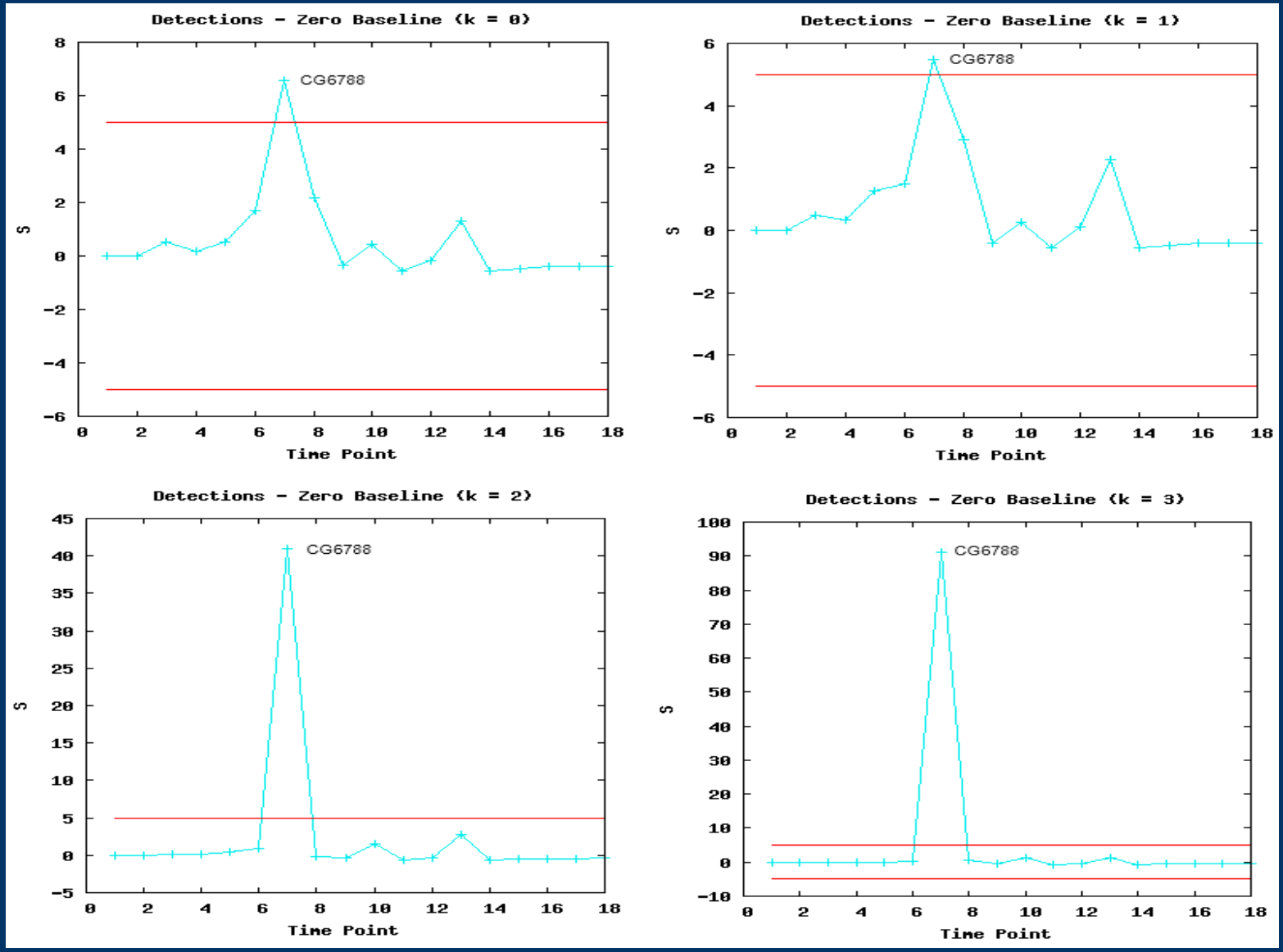
$$I\{\Psi_1(v) < \hat{\sigma}_{1,t,\tau}(v)c_c + \hat{\mu}_{1,t,\tau}(v)\}$$

Summarization of Anomaly Detection Methodology

- Reverse-engineer genetic regulatory network from the transcription profiling dataset.
- Generate co-expression graphs for each time point of the transcription profiling dataset.
- Run graph-based scan statistics algorithm against the time-series co-expression graphs.
- And you get...

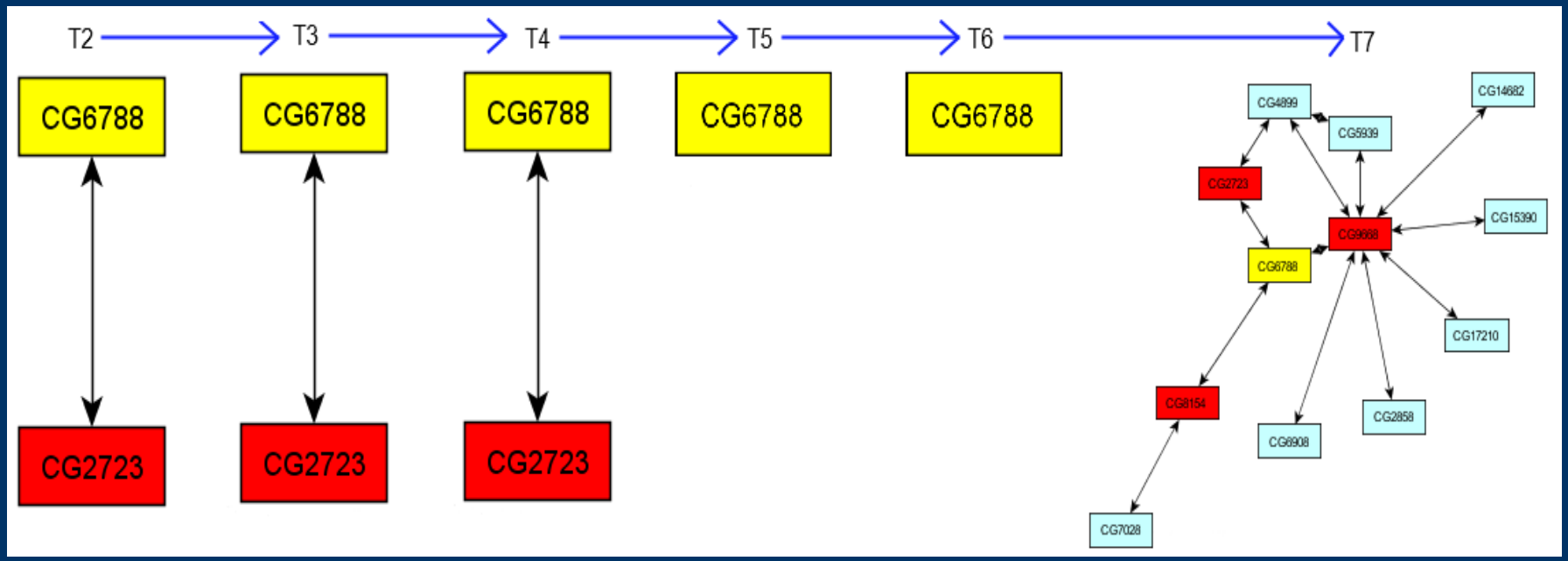
Anomaly Detections

Arbeitman Data – Zero Baseline



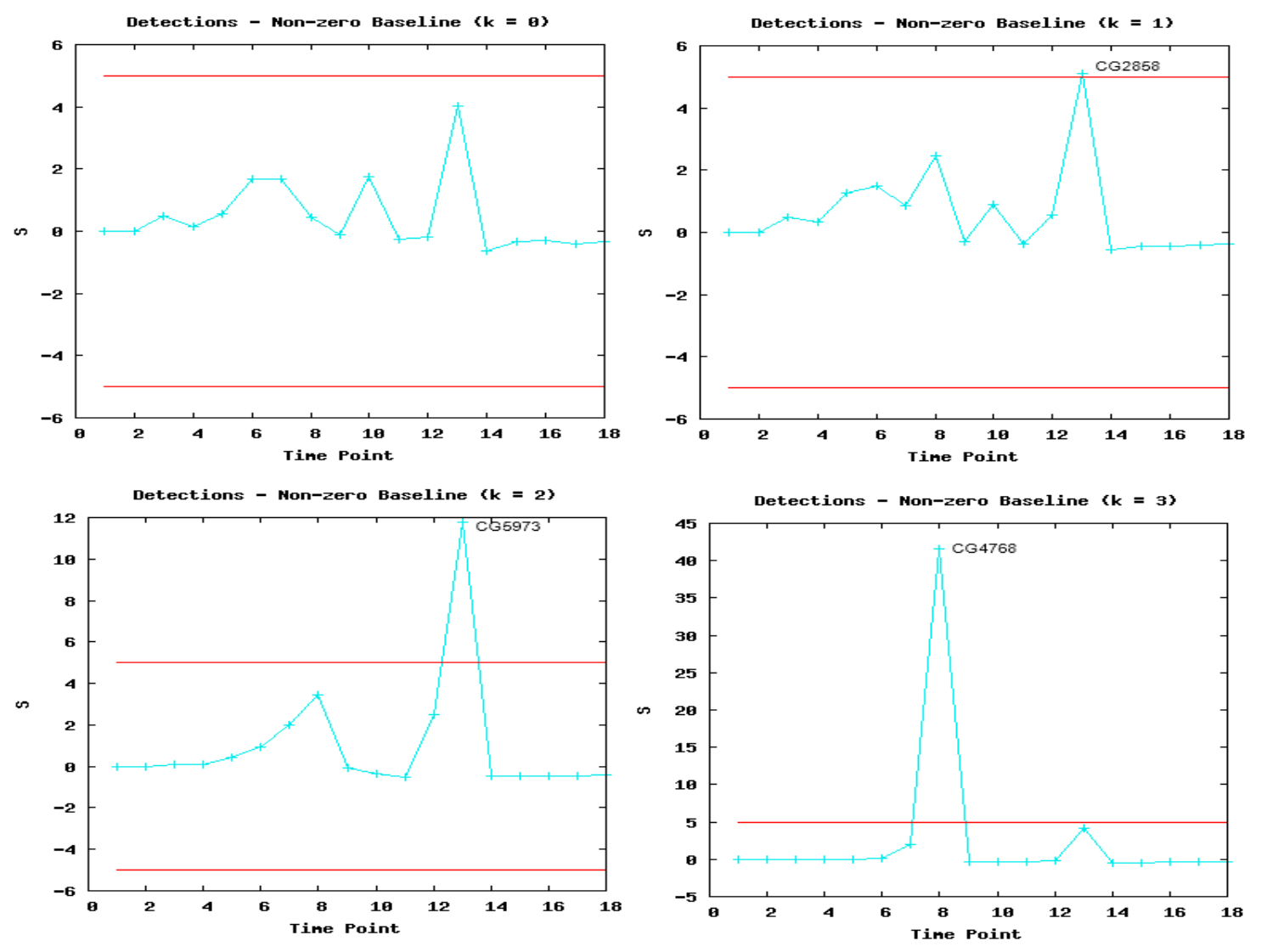
Anomaly Detections

Arbeitman Data – Zero Baseline Subraphs



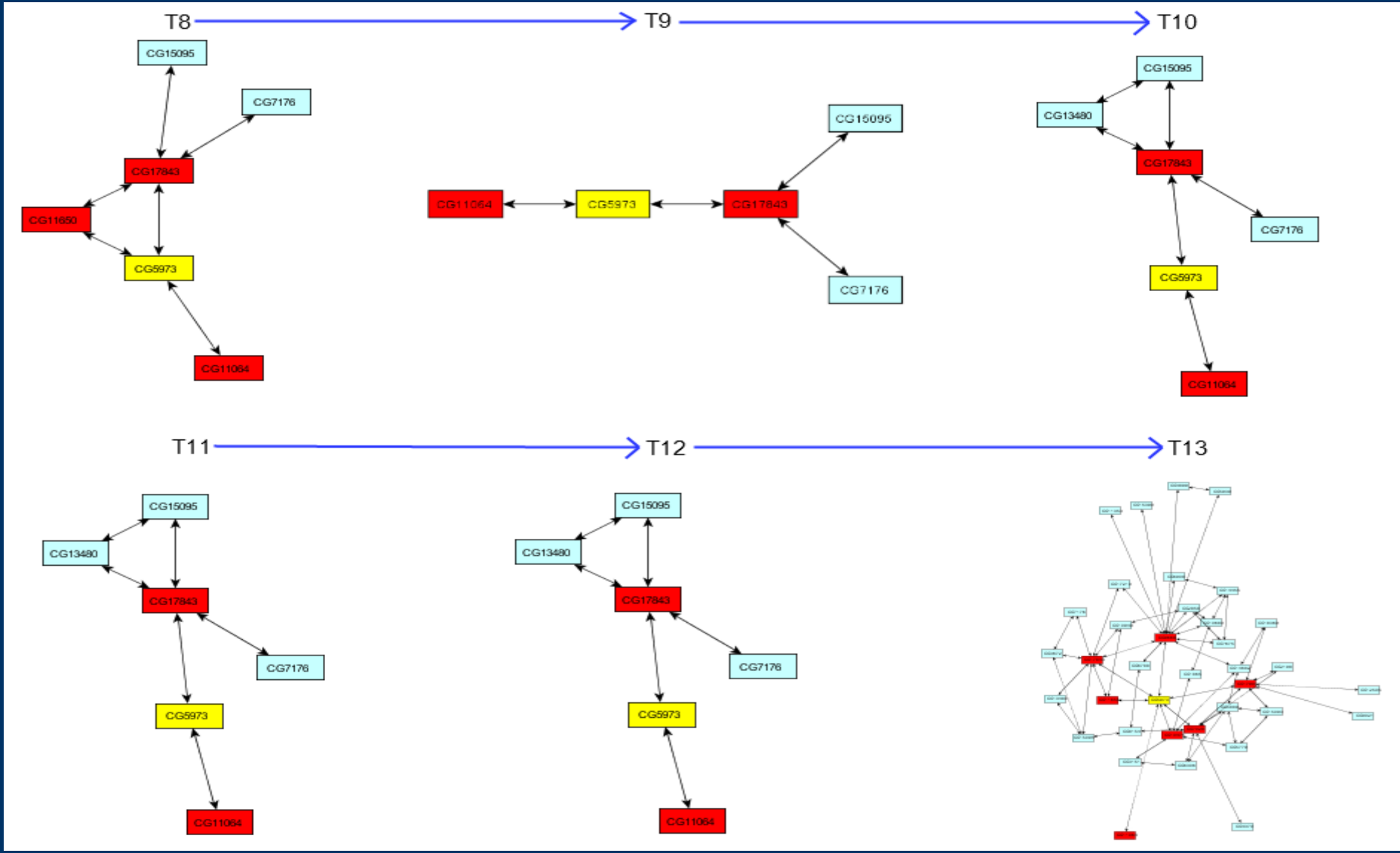
Anomaly Detections

Arbeitman Data – Non-Zero Baseline



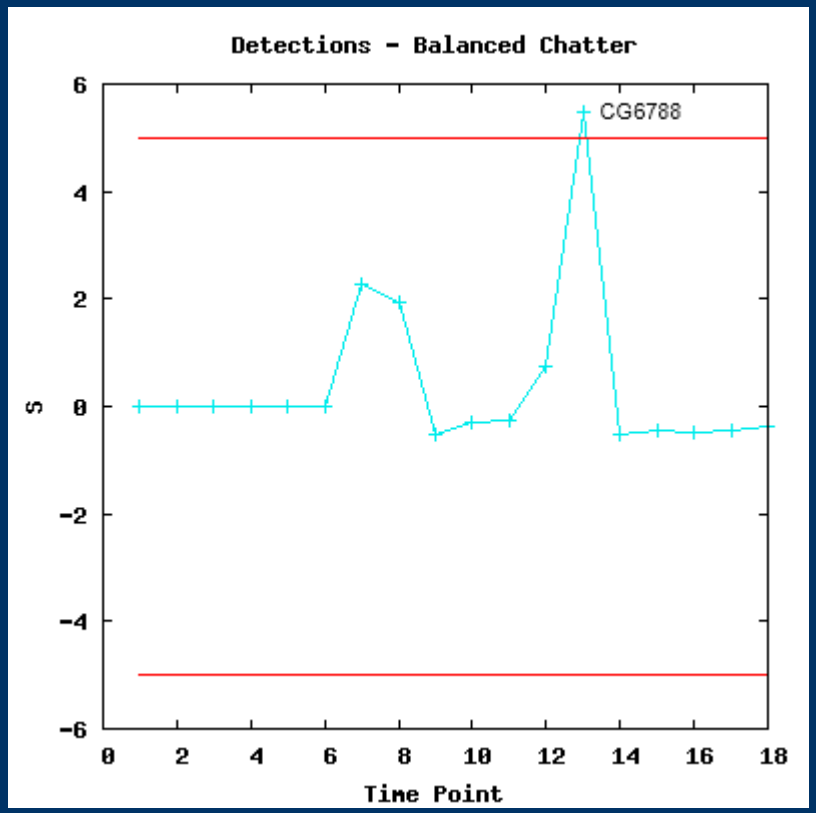
Anomaly Detections

Arbeitman Data – Non-Zero Baseline Subgraphs



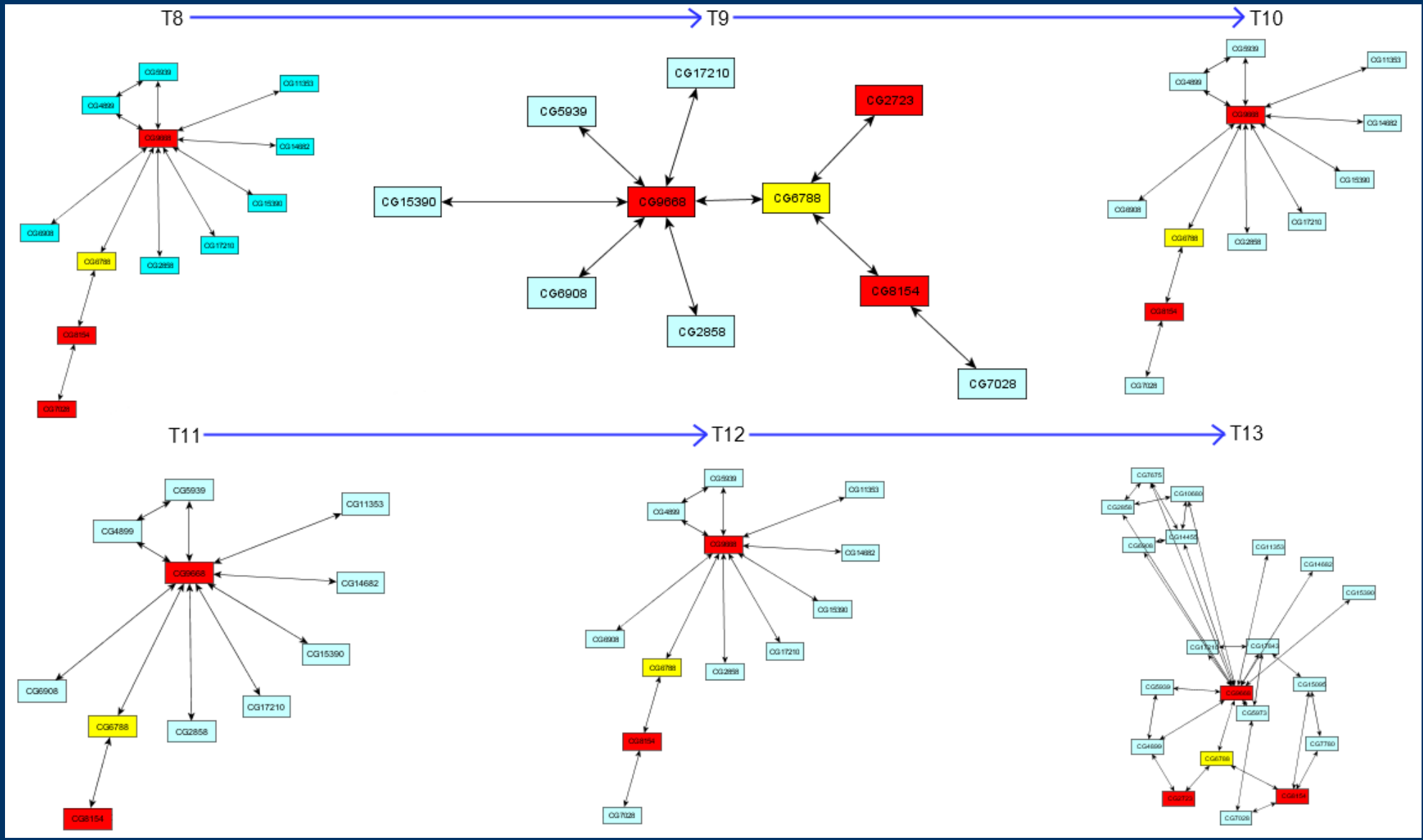
Anomaly Detections

Arbeitman Data – Balanced Chatter



Anomaly Detections

Arbeitman Data – Balanced Chatter Subgraphs



Summary

- Developed an anomaly detection technique for biological interaction networks, incorporating some of the dynamics of the system while using the simplifying network representation as the analysis framework.
- Successfully applied it to the pupal state time points of the Arbeitman Drosophila microarray dataset and detected several anomolous co-expression events.
- Although the characterization of static biological interaction networks and the interplay between them is far from complete, even for model organisms, we feel that it is nevertheless important to take the next logical step and begin to explore techniques for automated anomaly detection in these networks.

Future Work

- Provide visualization and analysis tools to aid researchers in the investigation of the detected anomalies
- Apply the algorithm to more datasets that have more time points.
- Apply the the algorithm to different types of biological data, such as metabolomic and proteomic data.
- Study the robustness and suitability of parameter choices to particular types of biological data.
- Use different graph invariants for the local statistic.
- Determine the biological significance of different graph invariants and indicator functions.

Thank You!