Using GOALIE to Analyze Time-course Expression Data and Reconstruct Kripke Structures

Marco Antoniotti
Department of Informatics, Systems and Communications
University of Milan Bicocca
ITALY

NYU CMACS NSF PI Meeting, New York, Oct 28-29 2010
Outline

• Interactions between experiments, data and interpretation
• Models of Biological Processes and Systems
  – Description (via controlled vocabularies and ontologies)
  – Reconstruction (via time-course analysis and statistical procedures)
  – Model Repositories
• Computational “Searches” for “models” (parameters, new interactions, etc)
  – Problems
    • Low sampling rate
    • Upsampling, optimization schemes
    • Models limitations
Analyzing Time-course Microarray Experiments

- Microrarray Experiments and Data
- “Enrichment” studies via Controlled Vocabularies and Ontologies (Gene Ontology and others)
- Model “reconstruction”
  - Similarity studies
  - Segmentation algorithms
  - Kernel methods
  - Results
- Future work

- Joint work with Bud Mishra, Courant NYU, Naren Ramakrishnan, Virginia Tech, Daniele Merico, University of Toronto, many others at NYU and UNIMIB
Microarray Experiments

- From laser scans readings, a numerical value corresponding to the relative expression of a “gene” is produced.

- When each raw data array scan corresponds to a given time-point under a specific condition, the final gene expression data matrix represents the temporal evolution of the gene expression.
Standard data-mining approaches to microarray data

• The results of microarray experiments have been studied by means of statistical techniques

• **Aim:**
  – To group together genes/probes that “behave similarly” under different experimental conditions (usually achieved by *clustering*)

• **Successful endeavor**
  – Several tools and libraries are provided to perform this kind of studies
  – Several publications produced with results in this field
  – Many of the studies reported still contain a considerable amount of “hand curation”
Standard data-mining approaches to microarray data

- The expression matrix is usually analyzed according to standard techniques:
  
  - Clustering enables to group together genes with a similar expression profile
  
  - Gene Ontology (GO) terms “Enrichment” enables to find statistically over-represented terms in given set of genes - i.e., clusters - thus providing some “functional” characterization
    - usually computed using some statistical significance test; e.g., Fisher’s exact test, Hypergeometric Test, Binomial Test, $\chi^2$ Test, plus various corrections
Gene Ontology (GO)

- GO is a controlled vocabulary for the functional annotation of genes
- GO is composed by three independent classifications, each of them having a hierarchical DAG structure
  - **MF**: Molecular Function (biochemical activity and molecule type)
  - **BP**: Biological Process
  - **CC**: Cellular Component

www.geneontology.org
Time-course microarray data

- Clustering is performed with all time-points together spanning the whole time-course

This amounts to assume that if genes are co-regulated across some time-points, they will also be co-regulated throughout the whole time-course

- However, co-regulation may be interrupted at a certain point
  - Different short-time and long-time response, e.g., DNA damage
  - Multiple-stages transcriptional program, e.g., development
GOALIE: a twist on “enrichment” studies

- GOALIE introduces a twist on enrichment studies by taking into account possible temporal variations of biological processes in time-course measurements.

- The key observation is that an “enrichment” of a set of genes/probes may vary depending on the length of the (time) vector of measurements.

- GOALIE assumes that the a time-course experiment has been broken down into windows and that each window has been clustered separately.

- Afterward the enrichment of each cluster in a window is compared with the enrichment of clusters in neighboring windows and all the possible relations are built in a DAG.
  - GOALIE provides several interfaces to explore, summarize and compare the DAGs pertaining to different experiments.
Piece-wise approach to time-course microarray data

- We split the time-course into discrete windows,
- Then compute clusters for each window separately,
- Finally reconnect clusters from adjacent windows exploiting similarity of Gene Ontology cluster enrichments.
Computational Modules

• In order to enhance the GOALIE software we concentrated on the components computational modules

• Computational modules are required for:
  1. Clustering (*Clique* [Shamir et al.], K-means, SVM, SOMs etc.; tool *Genesis* from TU-Graz and many other ones)
  2. Segmentation (PNAS 2010 [Ramakrishnan et al.])
  3. Gene Ontology (GO) enrichment (Fisher’s exact test etc.)
  4. Computing similarity among clusters from adjacent time-windows, based on GO enrichment (*ex-novo* – Kernel function)
  5. Select only relevant connections among clusters (*ex-novo*)

• In the rest of this presentation, the focus will be on the Kernel approach developed for module #4; #5 has been published in (CaOR 2010 [Antoniotti et al.])
Computing “Similarity” Using Graph Kernels

• The results of the first three steps of the algorithm consist in the “enrichment” of each cluster by a set of representative labels (GO terms)

• Next we want to see how similar two clusters are based on this labeling

• **Note**
  – This check may be useful to a biologist trying to track biological processes over time; e.g., trying to see which genes are involved in a certain process as time evolves
  – From a more abstract point of view this is a procedure that measures how two objects are similar
    • The similarity between the two objects is done in a **re-described** space (possibly with lower dimensionality)
    • In our case there is some more structure we want to exploit
Computing “Similarity” Using Graph Kernels

• Peculiarities of our method
  – Our objects are clusters ordered in a time-course
  – The labeling by GO terms does have a structure imposed by their hierarchical arrangement in a DAG

• Previous work
  – Similarity between objects of this kind is computed using various measures
  – In the specific case of labeling of gene sets, flat lists of symbols were used
    • Similarity computed Jaccard index
    \[ J(X,Y) = 1 - \frac{|X \cap Y|}{|X \cup Y|} \]

• Graph kernels can instead be used to take into account the DAG nature of the GO labels
  – Question: what is the performance of our Graph Kernel method w.r.t. a simple Jaccard index calculation?
Kernel Methods

When the existence of a non-linear pattern prevents from using a linear classification algorithm, the problem can be solved introducing a mapping function \( \Phi \) which projects the problem in a higher dimension space, where the pattern is linear.

\[ \Phi : R^N \rightarrow R^M \ (M > N) \]
Kernel methods

• How to perform the mapping?
  – We don’t really have to know the mapping $\Phi$ if we introduce a **Kernel function** $k$
    \[ k(x, y) = \langle \phi(x), \phi(y) \rangle_F \]
  – The internal product between the remapped points is compute by $k$ thus avoiding the explicit computation of $\Phi$ (the so called **Kernel Trick**)

• In order to be a proper Kernel, a function must be positive semi-definite and symmetric (Mercer’s Theorem)

• A Kernel function can also be used to induce a dissimilarity function (that’s exactly what we do)
A Kernel Function for Gene Ontology Graph Comparison

• Input: GO enrichment graph; i.e., sub-graphs of the overall GO taxonomy for each cluster
  – Each vertex is identified by a label - the GO term name - which is then used for walk matching
  – Each vertex has also an associated p-value label, from Fisher’s exact test, which is then used to compute a dissimilarity score between the walks
    • We work on GO sub-graphs (forests), obtained by filtering in only the terms with p-value < significance threshold
A Kernel Function for Gene Ontology Graph Comparison

- The computation (informally) proceeds in the following way
  1. We compute the (direct) graph product between the two GO sub-graphs
  2. We identify common walks in the product GO sub-graph
  3. We compute a weighted dissimilarity score for each walk
  4. We sum all the walk dissimilarities to get the total dissimilarity
A Kernel function for Gene Ontology graph comparison

• What are the advantages of our approach?
  – We explicitly take into account the hierarchical structure of GO cluster enrichments (Zoppis et al. 07 ISBRA)

• Next we concentrated on evaluating our approach
  – For a benchmark for our Kernel function we set up a comparison with a Jaccard Coefficient-based dissimilarity, working on GO enrichments as flat lists of terms
    • Once the dissimilarities are computed with both methods, we select only significant similarity patterns among clusters from adjacent windows (*)
  – We also consider a model manually curated by an expert
  – To quantitatively assess performance, we adopt the Loganantharaj et al (BMC Bioinformatics, 2006) **Total Cluster Cohesiveness** (TCC) score, which enables to assess the homogeneity of a cluster in terms of its GO terms; we compute TCC for groups of connected clusters (Merico et al. 07 KES-WIRN)

\[
\begin{align*}
\text{w1-c1} & \rightarrow \text{w2-c1} \\
\text{w1-c2} & \rightarrow \text{w2-c2} \\
\text{w1-c3} & \rightarrow \text{w2-c3}
\end{align*}
\]

\[
\text{TCC} = \frac{w1c1 + w2c1,2 + w1c3 + w2c3}{w1c1 + w2c1,2 + w1c3 + w2c3}
\]
GOALIE Interface

Clusters connection tree
Each level a “window”

Micro-array accessions

GO categories

Clusters information
Connection information
Cluster Information
GOALIE Interface

<table>
<thead>
<tr>
<th>Source cluster</th>
<th>GO Categories</th>
</tr>
</thead>
<tbody>
<tr>
<td>AA037229</td>
<td>GO:0007539 GO:0007160 GO:0007166 GO:0007165 GO:0005087</td>
</tr>
<tr>
<td>AA04326</td>
<td>GO:0006454 GO:0006796 GO:0007166 GO:0007175 GO:0001953</td>
</tr>
<tr>
<td>AA047257</td>
<td>NIL</td>
</tr>
<tr>
<td>AA083485</td>
<td>GO:0006412 GO:0009059 GO:0019538 GO:0009058 GO:000815</td>
</tr>
<tr>
<td>AA127160</td>
<td>NIL</td>
</tr>
<tr>
<td>AA128366</td>
<td>NIL</td>
</tr>
<tr>
<td>AA130633</td>
<td>NIL</td>
</tr>
<tr>
<td>AA147641</td>
<td>NIL</td>
</tr>
<tr>
<td>AA152237</td>
<td>GO:0003673 GO:0006950 GO:0050896 GO:0008152 GO:000758</td>
</tr>
<tr>
<td>AA187349</td>
<td>GO:0008202 GO:0006766 GO:0006629 GO:0006091 GO:000611</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Destination cluster</th>
<th>GO Categories</th>
</tr>
</thead>
<tbody>
<tr>
<td>GO:0006873</td>
<td>cell ion homeostasis</td>
</tr>
<tr>
<td>GO:0007160</td>
<td>cell-matrix adhesion</td>
</tr>
<tr>
<td>GO:0007599</td>
<td>blood coagulation</td>
</tr>
<tr>
<td>GO:0019275</td>
<td>cell homeostasis</td>
</tr>
<tr>
<td>GO:0042592</td>
<td>cell homeostasis</td>
</tr>
<tr>
<td>GO:0005081</td>
<td>ion homeostasis</td>
</tr>
<tr>
<td>GO:00060817</td>
<td>coagulation</td>
</tr>
<tr>
<td>GO:0006078</td>
<td>regulation of cell migration</td>
</tr>
</tbody>
</table>

GO categories sharing with "destination" cluster but not "source" cluster:
- GO:0006873: cell ion homeostasis
- GO:0007160: cell-matrix adhesion
- GO:0007599: blood coagulation

GO categories describing "source" cluster but not "destination" cluster:
- GO:0001953: skeletal development
- GO:0006451: energy pathways
- GO:0006796: protein complex assembly
- GO:0006766: vitamin metabolism
- GO:0008202: negative regulation of cell migration
- GO:0006091: nucleotide metabolism
- GO:0006078: regulation of cell migration
GOALIE Interface
## GOALIE Interface

### Summary Comparison View of Two Cell Cycle Experiments

<table>
<thead>
<tr>
<th>Spellman Yeast Cell Cycle Data - Elutriation</th>
<th>Spellman Yeast Cell Cycle Data - Cdc15</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>protein amino acid acetylation</strong></td>
<td><strong>protein amino acid acetylation</strong></td>
</tr>
<tr>
<td><strong>microtubule nucleation</strong></td>
<td><strong>microtubule nucleation</strong></td>
</tr>
<tr>
<td><strong>mitotic anaphase</strong></td>
<td><strong>mitotic anaphase</strong></td>
</tr>
<tr>
<td><strong>cytochrome complex assembly</strong></td>
<td><strong>cytochrome complex assembly</strong></td>
</tr>
<tr>
<td><strong>carnitine metabolism</strong></td>
<td><strong>carnitine metabolism</strong></td>
</tr>
<tr>
<td><strong>response to external stimulus</strong></td>
<td><strong>response to external stimulus</strong></td>
</tr>
<tr>
<td><strong>aspartate family amino acid biosynthesis</strong></td>
<td><strong>aspartate family amino acid biosynthesis</strong></td>
</tr>
<tr>
<td><strong>vitamin or cofactor transport</strong></td>
<td><strong>vitamin or cofactor transport</strong></td>
</tr>
<tr>
<td><strong>glycerol biosynthesis</strong></td>
<td><strong>glycerol biosynthesis</strong></td>
</tr>
<tr>
<td><strong>polyl biosynthesis</strong></td>
<td><strong>polyl biosynthesis</strong></td>
</tr>
<tr>
<td><strong>S phase of mitotic cell cycle</strong></td>
<td><strong>S phase of mitotic cell cycle</strong></td>
</tr>
<tr>
<td><strong>regulation of translation</strong></td>
<td><strong>regulation of translation</strong></td>
</tr>
<tr>
<td><strong>septin assembly and septum formation</strong></td>
<td><strong>septin assembly and septum formation</strong></td>
</tr>
<tr>
<td><strong>urea cycle intermediate metabolism</strong></td>
<td><strong>urea cycle intermediate metabolism</strong></td>
</tr>
<tr>
<td><strong>response to pheromone</strong></td>
<td><strong>response to pheromone</strong></td>
</tr>
</tbody>
</table>
Yeast Cell Cycle benchmark

- Cell Cycle is a multi-stage phenomenon (phases), therefore co-regulation patterns may change across time
  - In [Ramakrishnan et al. 2010] we consider different datasets regarding YCC and Yeast Metabolic Cycle
  - In particular, we consider two windows: G1>S and G2>M>G1
- We use Spellman microarray yeast cell cycle data (1998; a well known benchmark for testing novel analysis tools and methods)
  - CDC15-mutant synchronization
  - ALPHA factor synchronization
Comparison results using KL segmentation

Yeast “Metabolic” Cycle Segmentation Comparison: 8 segments inferred
Results

Black solid lines represent connections found both by the manual and automatic methods; Bold lines represent the strongest connections. Black dashed lines represent connections found only by the manual method. Grey dash-dotted lines represent connections found only by the automatic methods..
Results

Results overview

• Main results were generated for Alpha subset (2 windows), displaying a substantial convergence between the three methods
  – Numerical results are comparable with Jaccard method
  – Kernel method is more “correct” from the information point of view
  – Kernel method is more computationally intensive

• Preliminary results were also generated for CDC15 subset, displaying a better performance of Kernel over Jaccard

Results (Alpha subset)

<table>
<thead>
<tr>
<th>Distance</th>
<th>TCC</th>
<th>threshold</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jaccard</td>
<td>94.28</td>
<td>0.05</td>
</tr>
<tr>
<td>Jaccard</td>
<td>92.95</td>
<td>0.01</td>
</tr>
<tr>
<td>Jaccard</td>
<td>92.95</td>
<td>0.005</td>
</tr>
<tr>
<td>Kernel</td>
<td>92.95</td>
<td>0.01</td>
</tr>
<tr>
<td>Kernel</td>
<td>94.63</td>
<td>0.05</td>
</tr>
<tr>
<td>Manual</td>
<td>92.27</td>
<td>N/A</td>
</tr>
</tbody>
</table>
Problems

• **Low sampling rate**: biological experiments usually have a way too low sampling rate
  – **Ok** for long term observations at equilibrium
  – **Not ok** for transients and discontinuities detection
    • **Assumption**: transients and discontinuities are interesting

• **Solutions**
  – **Upsampling** after fitting the data to a set of interpolating functions (rational functions or polynomials)
  – **Merging** of different data sources
    • Several institutions and databanks (e.g., GEO) contain several experiments
    • “Related” experiments can be combined to yield a Virtual Time-Course Experiment that organized the extant corpus of knowledge
Current and future research

• Connection ordering between clusters
  – Method based on optimization of (average) entropy orders connections according to a decrease in the uncertainty of the result graph Kernel similarity between the labeling of two clusters (Antoniotti et al. CaOR 2010)
    • “Complementary” with work on segmentan based on KL divergence published in Ramakrishnan et al. PNAS 2010

• Sample classification (i.e. VTE reconstruction) can be performed if there is an appropriate model of the underlying biological system
  – Ontology research
    • Signs Symptoms Findings Workshop in Milan, 3-4 September 2009
Current and future research

• **Temporal Series Reconstruction** is a hard problem (deterministically akin to the Traveling Salesman Problem)
  – Bar-Joseph models based on EM optimization procedure
  – Magwene and Kim procedure based on heuristic MST built on top of PQ-trees
  – Lack of data points is a problem

• **Prediction Models**
  – What happens if we “extend” a time course in the future?
Acknowledgements

- BiMiB Lab, Dipartimento Informatica Sistemistica Comunicazione Milano-Bicocca bimib.disco.unimib.it
- Courant Bioinformatics Group New York University
  - S. Kleinberg, A. Sundstrom, A. Witzel, S. Paxia, B. Mishra
- Virginia Tech
  - S. Tadepalli, N. Ramakrishnan
- IFOM, Milan
  - M. Gariboldi, J. Reid, M. Pierotti
- Bader Lab, Donnelly Centre for Cellular and Biomolecular Research, University of Toronto
  - G. Bader, D. Merico
- Virtual Physiological Human Network of Excellence, European Commission FP7
- Regione Lombardia
- National Science Foundation EMT Program
- European Commission Marie Curie Program FP6
Thank you!