

Learning Predictive Models of Gene Dynamics

Nick Jones, Sumeet Agarwal



An Aim of Systems Biology

- How about predicting what cells are going to do?
 - Quantitatively predict the response of cells to multiple new inputs
- Prediction = understanding?
- In biomedicine this matters less.

The brains of cells

- For getting a grip on the slower responses of cells (development/environmental change) it's TF nets.
- Big academic industry of trying to deduce Transcription Factor networks using a variety of tools.
- Old Review: *Genome Res. 2006 Unraveling transcription regulatory networks by protein-DNA and protein-protein interaction mapping.* Walhout
- DREAM project devoted to obtaining nets

A Predictive Model for Transcriptional Control of Physiology in a Free Living Cell

- Bonneau et al Cell '07

SUMMARY

The environment significantly influences the dynamic expression and assembly of all components encoded in the genome of an organism into functional biological networks. We have constructed a model for this process in *Halobacterium salinarum* NRC-1 through the data-driven discovery of regulatory and functional interrelationships among ~80% of its genes and key abiotic factors in its hypersaline environment. Using relative changes in 72 transcription factors and 9 environmental factors (EFs) this model accurately predicts dynamic transcriptional responses of all these genes in 147 newly collected experiments representing completely novel genetic backgrounds and environments—suggesting a remarkable degree of network completeness. Using this model we have constructed and tested hypotheses critical to this organism's interaction with its changing hypersaline environment. This study supports the claim that the high degree of connectivity within biological and EF networks will enable the construction of similar models for any organism from relatively modest numbers of experiments.

The Pipeline

- 1) Get a handle on key genes (sequence and annotate)
- 2),3) Kick (KO's and environment) the system in well motivated ways and measure transcripts (possibly through time)
- 3) Cluster the Genes using black magic 1.
- 4) Train a system that associates Transcription Factor levels and Environmental state with the ensuing average levels of clustered genes (black magic 2).
- 5) Recycle and observe to taste
- 6) Make sure that you test it out on some new data at the end.

Two key refs for the following

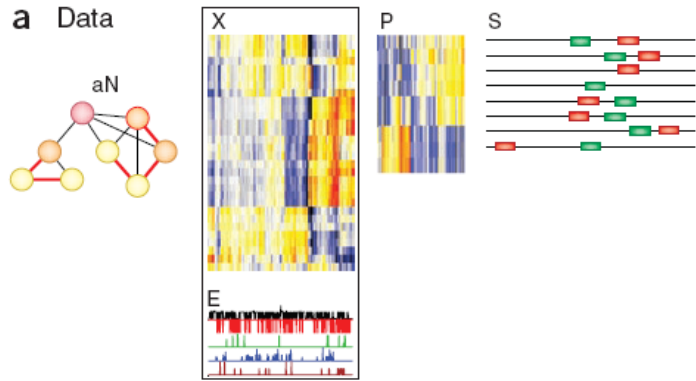
- The authors wouldn't claim to be best – we're discussing these because they fit into an interesting pipeline.
- BMC Bioinformatics '06 *Integrated biclustering of heterogeneous genome-wide datasets for the inference of global regulatory networks*
- Genome Biology '06 *The Inferelator: an algorithm for learning parsimonious regulatory networks from systems-biology data sets de novo*



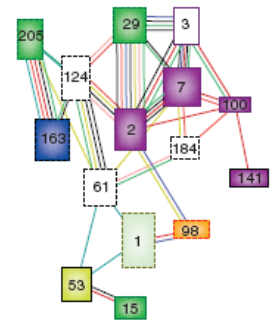
(fairly) New bits

- They integrate multiple data sources in their biclustering
- They perform network inference that is dynamic and efficient

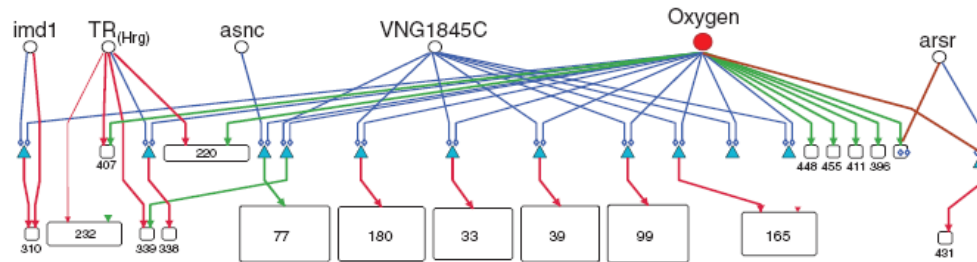
a Data



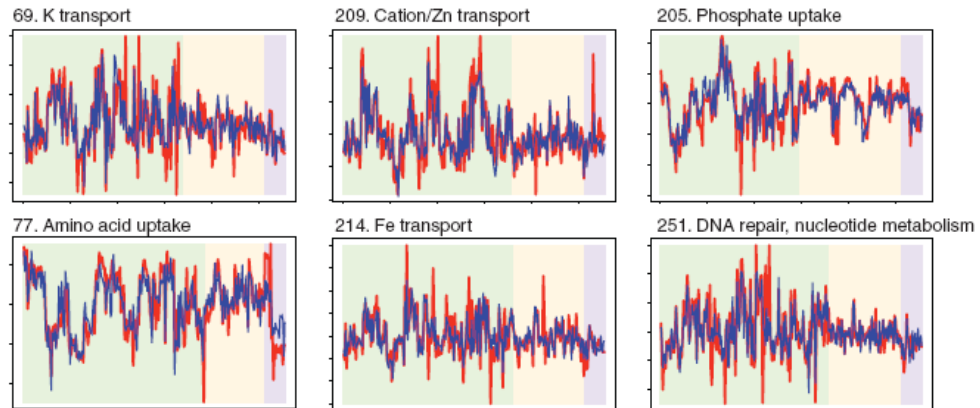
b (Bi)clustering





c Dynamical network model



d Prediction



- 
- 
- What is clustering?
 - Why cluster?
 - What is a bicluster? “a a subset of genes that exhibit compatible expression patterns over a subset of conditions”
 - Why bicluster?
 - Review (oldish): Bioinformatics 2006 *A systematic comparison and evaluation of biclustering methods for gene expression data*

Integrated Biclustering


- Aim is to create biclusters which combine information from
 - A) Gene expression data in various conditions
 - B) Various networks of gene associations (metabolic, associations, PINs)
 - C) Upstream, regulatory, sequences
 - (kitchen sink included in the above)
- Departure from expression only biclustering

Dynamical Model to Fit

$$\tau \frac{dy}{dt} = -y + g(\beta \cdot Z)$$

- *Average* transcript level of a bicluster is y
- τ specifies relaxation timescale
- $x(t)$ is the vector of TF mRNA concentrations and environmental states at time t : $x = (\dots, [TF]_p, \dots, e_q, \dots)$
- $Z = (\dots, \min(x_l, x_m), \dots)$ for all l, m ($l \geq m$).

$$g(\beta Z) = \begin{cases} \beta Z : & \text{if } \min(y) < \beta Z < \max(y) \\ \max(y) : & \text{if } \beta Z > \max(y) \\ \min(y) : & \text{if } \beta Z < \min(y) \end{cases}$$

- 
- Aim: For each bicluster (with average concentration y) find a β st for all $x(t)$ we make the best predictions for changes in the average concentration of the bicluster: Δy .
 - We make no attempt to predict how $x(t)$ evolves (except indirectly). So it's not a fully coupled system of equations where every entity has an equation of motion.

- Ordinary least squares estimate for β :

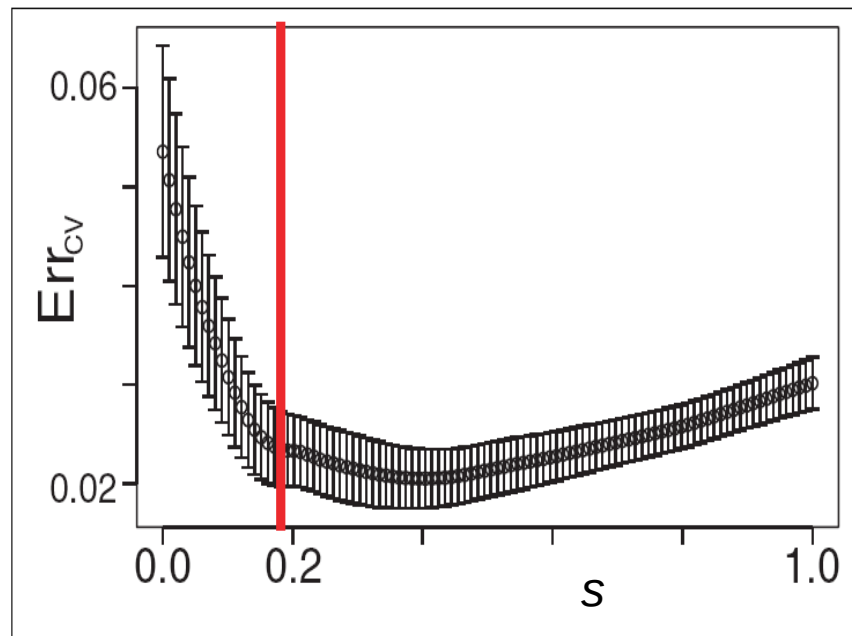
$$\epsilon_i(\beta) = \min_{\beta} \left[\sum_{k=1}^{T-1} \left| \tau \frac{y_i(t_{k+1}) - y_i(t_k)}{t_{k+1} - t_k} + y_i(t_k) - g(\beta \cdot Z_k) \right|^2 \right]$$


- Where we consider bicluster i and ϵ is our error for the best choice of β and we have T time steps that we average our error over


- Enforce model parsimony

$$|\beta|_1 \leq s|\beta_{ols}|_1$$

- Vary s with a balance of minimizing the Cross Validation Error and having s as small as possible



- 
- But...this is all still too messy...actually they first find a β and then find the top 5 single factors and top 2 pairwise interactions.
 - Alternative of choosing 7 non-zero entries out of a vector of 88 environments and TF's, plus all possible pairwise interactions, leads to $\sim 10^{21}$ choices

- 
- Then they go on and make some impressive predictions about the change in average transcript levels of clusters under
 - New combinations of pre-trained environments
 - New environments
 - New combinations of Transcription Factors and environments by Knock Outs
 - Obtain a 0.8 correlation between predicted and measured mRNA levels

Questions

- How does this differ from a big look-up table?
- How useful is it to know the time evolution of arbitrary clusters?
- Mightn't we expect the state of one cluster to affect the state of another?
- If I exhaustively train my system then do I have much scope for new behaviour – especially if I don't use radically new inputs?
- How sensitive is this to the choice of biclustering?
- How can we know whether this predictive ability will be preserved in new circumstances?
- Can we not try and predict the TF dynamics also?

Qualitative Modelling

- We can define qualitative relations between variables and attempt to learn a model based on these
- The variables themselves become qualitative: e.g., instead of tracking the actual expression level of a gene, we just represent it by a certain number of discrete states, say ON and OFF
- Relational learning techniques like Inductive Logic Programming can be used to infer such models from data

Example

- Suppose we have a gene whose expression level Y is regulated by an activator (level $X1$) and a repressor (level $X2$)

- Then a qualitative model for it might be:

DERIV(Y , DY)	// DY is the derivative
MPLUS($X1$, Prod Y)	// Production is incr. fn.
MMINUS($X2$, Prod Y)	// Decreasing fn. of $X2$
MPLUS(Y , Deg Y)	// Degradation rate
ADD(DY , Deg Y , Incr Y)	// Net change is sum



Advantages and Challenges

- Qualitative models are one way of dealing with highly noisy expression data sets, by abstracting away the precise measurements
- Have to come up with an appropriate discretisation of variables
- This approach has worked well for small-scale models, but will it scale to thousands of genes? Do we have enough data?

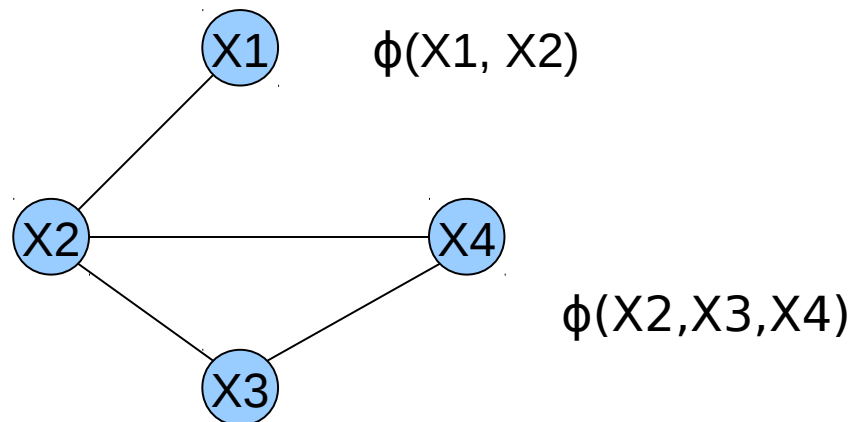


Probabilistic Models

- Another approach is to attempt to model joint probability distributions over gene expression levels
- Since the full joint distribution over thousands of genes will be not be learnable from realistically-sized datasets, we need to partition it in some way
- One way of doing this is to use a Markov Random Field (MRF) model

MRFs

- We define a graph of linkages/correlations between different genes, based on domain knowledge
- The graph is partitioned into “components”, and a distribution function is learnt independently for each component



Summary

- Outlined a pipeline for a predictive cellular biology, based on a machine learning approach to infer gene (cluster) expression dynamics
- Looked at qualitative modelling as a possible alternative to get around issues of noise and provide a natural framework for logical relations (AND, OR...)
- Probabilistic modelling using Random Fields is another possible approach, provided we can first extract sufficient domain knowledge