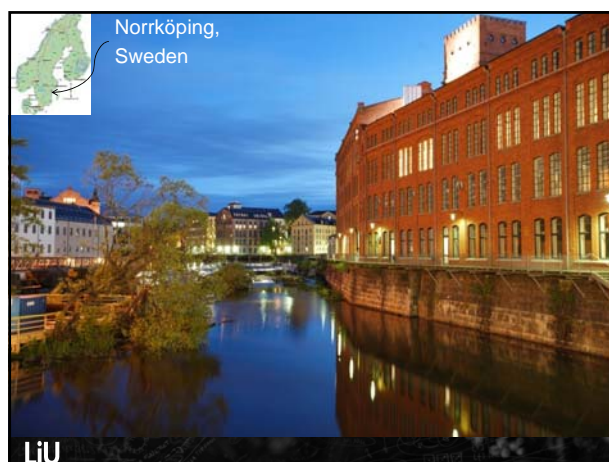


# Integrated Inference and Systems Analysis of Gene Regulatory Networks Based on High-throughput Data

Michael Hörnquist  
Linköping University, Dept. of Science and Technology

LiU expanding reality



## Presentation today

- Mission
- Group presentation
- Bragging
- Examples of recent work
  - **Large-scale inference by soft integration and the elastic net**
  - **System Analysis of *Linearized* GRNs**
- Final remarks

LiU

Please interrupt me with questions.

It is a privilege to speak *with*,  
rather than to speak *to*.

LiU

## Our mission

- Develop useful algorithms for mining knowledge out of high-throughput data from molecular biology



### Including

- Inference and analysis of biological networks
- Prediction of experiment
- Impact on the future development of computational medicine

LiU

## Group and collaborators

At LiU in Norrköping	Collaborators today
<ul style="list-style-type: none"> <li>▪ Michael Hörnquist</li> <li>▪ Mika Gustafsson, PhD-student</li> <li>▪ Anna Lombardi</li> </ul>	<ul style="list-style-type: none"> <li>▪ Jesper Tegnér, KI (LiU)</li> <li>▪ Johan Björkegren, KI</li> <li>▪ Shohreh Maleki, KI</li> <li>▪ Vlad Bajic, KAUST</li> <li>▪ Harukazu Suzuki, RIKEN</li> </ul>

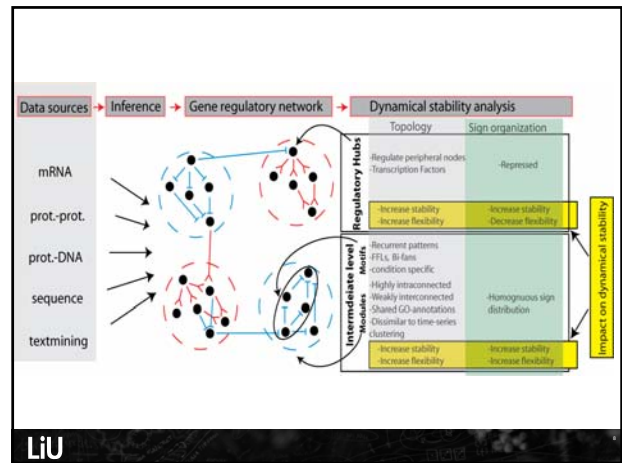
LiU

### Some recent highlights

- DREAM2, prediction of in-silico network, 2007
  - Best performance algorithm
- DREAM3, prediction of gene expression, 2008
  - Best performance algorithm



LiU



LiU

### First part

Large-scale inference by soft integration and the elastic net

(essentially our contribution to DREAM3)

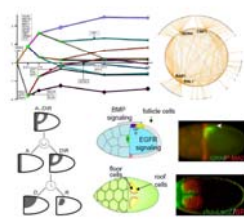
LiU

### On the DREAM3 challenges

DREAM = Dialogue on Reverse Engineering and Assessment Methods

Abstracts of papers, posters and talks presented at the 2008 Joint RECONB Satellite Conference on REGULATORY GENOMICS - SYSTEMS BIOLOGY - DREAM3

1. Signaling Cascade Identification - Infer a signaling network from flow cytometry data
2. Signaling Response Prediction - Predict missing protein concentrations from a large corpus of measurements
3. Gene Expression Prediction - Predict missing gene expression measurements
4. In Silico Network Challenge - Infer simulated gene regulation networks



LiU

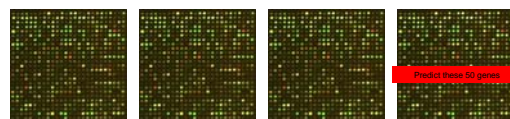
### Why are challenges like DREAM important?

- The assessment of algorithms in systems biology is difficult
- Each outcome, e.g., each edge in an inferred network, is de facto a prediction
- Validation experiment *after* the prediction (normal science):
  - Wet-lab experiments can confirm individual edges or levels, but not genome-wide predictions (otherwise would much of systems biology be unnecessary)
  - It is seldom stated how many experiments have been carried out before the desired results are reached
- Validation experiment *before* the prediction (astronomy?):
  - When the results are known beforehand, there is a risk of fine tuning and over fitting

LiU

### On the Gene Expression Challenge

- Gene expression time course data is provided for four different strains of yeast (*S. Cerevisiae*), after perturbation of the cells.
- Each time course comprises 9335 probes at 8 different times
- Predict the rank order of induction/repression of a subset of 50 genes (the "prediction targets") in one of the four strains for each of the 8 times.
- You are allowed to use any information that is in the public domain **Integrative network inference**



LiU

### Going through the math

Basic equation  $\hat{x}_i(t) = a'_i + \sum_j w'_{ij} x_j(t)$

Turned inside out  $x_i(t) = a_i + \sum_{j \in T} \tilde{w}_{ij} \hat{x}_j(t) + \sum_{j \in T} w_{ij} x_j(t)$

Exploring all perfect fits (much more parameters than experiments), with minimal L1- and L2-norms of the coefficients, respectively, yields **no derivatives**.

LiU

### Going through the math

~~Regularized least absolute deviation (RLAD)~~

~~$\sum_k \left| x_i(t_k) - a_i - \sum_{j \in T} w_{ij} x_j(t_k) \right| + \lambda_i \sum_j |w_{ij}|$~~

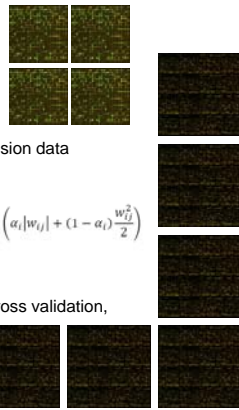
Least squares with the elastic net

$\sum_k \left( x_i(t_k) - a_i - \sum_{j \in T} w_{ij} x_j(t_k) \right)^2 + \lambda_i \sum_j \left( \alpha_i |w_{ij}| + (1 - \alpha_i) \frac{w_{ij}^2}{2} \right)$

Cross-validation gives the best performance for least squares with the elastic net

LiU

### Going through the math, more expression data



Give different weights to different expression data

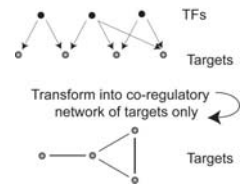
$\sum_k v_k \left( x_i(t_k) - a_i - \sum_{j \in T} w_{ij} x_j(t_k) \right)^2 + \lambda_i \sum_j \left( \alpha_i |w_{ij}| + (1 - \alpha_i) \frac{w_{ij}^2}{2} \right)$

Let the predicted power, as given by cross validation, determine the weights

LiU

### Going through the math, TF-DNA binding data

- TF:s expressed on low levels
- Often they need activation
- Concentrate on co-regulations



Weighted shared fraction of TF:s

$M_{ij} = \frac{\sum_n TF_{in}^{(e)} TF_{jn}^{(e)}}{\sqrt{(\sum_n TF_{in}^{(e)}) (\sum_n TF_{jn}^{(e)})}} + \frac{\sum_n TF_{in}^{(e+p)} TF_{jn}^{(e+p)}}{\sqrt{(\sum_n TF_{in}^{(e+p)}) (\sum_n TF_{jn}^{(e+p)})}}$

LiU

### Going through the math, the full model

Eventually, the objective function looks like

$\sum_k v_k \left( x_i(t_k) - a_i - \sum_{j \in T} w_{ij} x_j(t_k) \right)^2 + \lambda_i \sum_j \frac{1}{1 + \beta M_{ij}} \left( \alpha_i |w_{ij}| + (1 - \alpha_i) \frac{w_{ij}^2}{2} \right)$

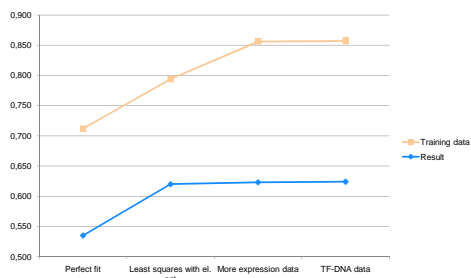
We determine 207 103 parameters, by cross-validation for  $\alpha, v_k, \beta, \lambda_i$  using R-package glmnet (Friedman et al.)

LiU

And the result is...

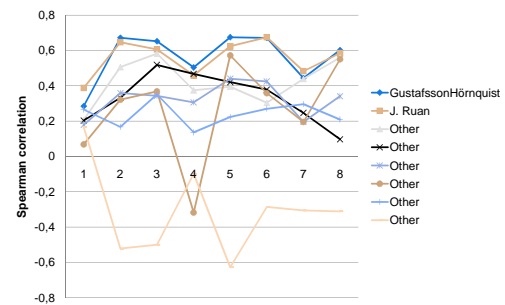
LiU

### Results during the development



LiU

### Results per experiment for each participant



LiU

### What did Ruan do?

- Standard KNN-model (predicts value as the mean values of the  $k$  nearest neighbours)
- Distance measured by Euclidean (L2) measure
- Cross validation yielded  $k=10$ , global parameter
- No external data, no prior knowledge
- How could he succeed so well?

LiU

To summarize...

LiU

### Conclusions, integrated inference

- Possible to predict gene levels from high-throughput data
- The elastic net can handle correlated variables and do subset selection
- Possible to combine expression data from different experiments
- Soft fusion of expression data with structural data can enhance the performance
- Lot of possible improvements, since an algorithm with no external data sources performed *almost* as well as ours

LiU

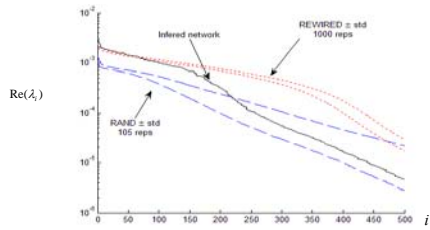
### Second part

Systems Analysis  
of  
*Linearized Gene Regulatory Networks*

LiU



### Empirical findings



- Leading eigenvalue of the inferred network is unexpectedly large
- Most of the positive eigenvalues are unexpectedly small
- Unstable system in the linear systems sense

LiU

### Consequences on dynamical stability

- In a linear system, a positive real part of any eigenvalue implies an exponential growth of the corresponding eigenmode
- In a linearized system, a positive real part reflects a rapid movement from the working point were the linearization is valid
- In biological systems, it enables switches with a short response time for selective signals
- Must be compensated by non-linear effects
- Too many instabilities, may lead to an ever diffusion from the working point

LiU

### Quantifying stability and flexibility

- $I$  is positively correlated to the rate which with the system drifts due to random fluctuations
- $S$  measure the degree of stability:

$$I = \sum_{j=1}^{N_i} \text{Re}(\lambda_j),$$

$$S = 1 - \frac{I}{I_{MAX}}$$

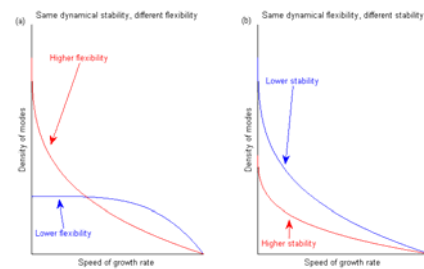
- $PR$  is a continuous measure of how many positive eigenvalues are significant larger than the rest
- $F$  is between zero and unity

$$PR = \frac{\left( \sum_{j=1}^v [\text{Re}(\lambda_j)]^2 \right)}{\sum_{j=1}^{N_i} (\text{Re}(\lambda_j))^2},$$

$$F = \frac{N_i - PR}{N_i - 1}$$

LiU

### Visualizing the concepts



LiU

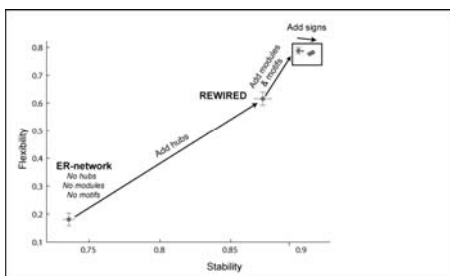
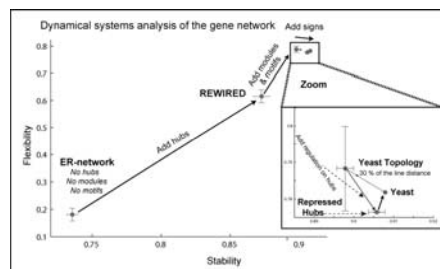


Figure 4

LiU



LiU

## Speculations on the reasons

- Hubs induce stability
  - Its presence also coincides with several non-regulators, which imply several marginally stable eigenmodes
- Hubs induce flexibility
  - Perturbation on any of the regulatory hubs is likely to propagate downstream and due to the complex structure grow
- Hub repression increases stability, but decreases flexibility
  - This is an effect of decreasing the real parts of the largest eigenmodes, from a negative in-regulation of regulatory hubs
- Higher order organization increases flexibility most, but also stability
  - For example modules might be well isolated sub-networks triggered by some specific signal, but stable against most perturbations (a miniature of the characteristics of the whole system)

LiU

## Summary, systems analysis of GRNs

- Refined concepts enable genome-wide systems analysis of ODE models around a working point
- Might be incorporated to guide the inference process
- Topology and dynamics jointly enables *stable yet flexible* systems

LiU

At the end of the day, what do we have?

- High-throughput data provides a challenge to analyze, but it is possible (for those still hesitating)
- Integration of several data sources can guide an inference process, but is highly non-trivial
- Concepts such as stability and flexibility can increase our understanding of the subtle interplay between topology and dynamics

LiU

LiU

Thank you for listening

[www.itn.liu.se/~micho](http://www.itn.liu.se/~micho)  
 michael.hornquist@liu.se