COMPUTATIONAL MODELING OF TRANSCRIPTION IN YEAST

David R. Haynor
University of Washington/Rosetta Inpharmatics

February 25, 2003

Outline

Motivation

Model

Computational aspects

Validation

Future work

The problem.

Given a set of connections between transcription factors and genes (possibly including some false positives), determine the strength of the interactions between the factors and the genes, and the variation in effective transcription factor levels from experiment to experiment.

Motivation

Transcription factor activation of genes represents the final endpoint of signaling pathways. TF levels are hard to measure directly.

Partitional clustering does not accurately represent underlying transcriptional control.

Better understanding of the strength of TF-gene interactions should improve *in silico* predictions of regulation.

A model for yeast transcription

$$\log(Y_{eg}) = \sum_{f \in F(g)} \alpha_{fg} X_{ef} + n_{eg}$$

Notes: 1. The noise variance may depend on *e,g*.

2. Missing data is handled by data weighting.

3. X_{ef} is the "effective" level of factor *f* in expt. *e*, $X \sim 0$ at baseline.

4. The association of *g* with *F(g)* is the *connection data*, and is assumed to be given.

5. Model may be better for metabolic/signaling pathways than for developmental programs.

Computational aspects

Identifiability: normalize RMS value of X_{ef} to be 1.

Minimization: alternately fix X's and α's and solve the resulting series of linear weighted-least squares problems.

Polish with minimizer suitable for large numbers of variables (10-200K) such as LBFGSB.

Time for G=2600, E = 725, F = 140: < 2 mins.

Data(1)

Publicly available data sources from yMGV links

Transformed to ratio data; weight = 0/1 for bad/good data (9% bad).

725 experiments x 6316 ORF's

Median normalization in *e* and *g*

Datasets

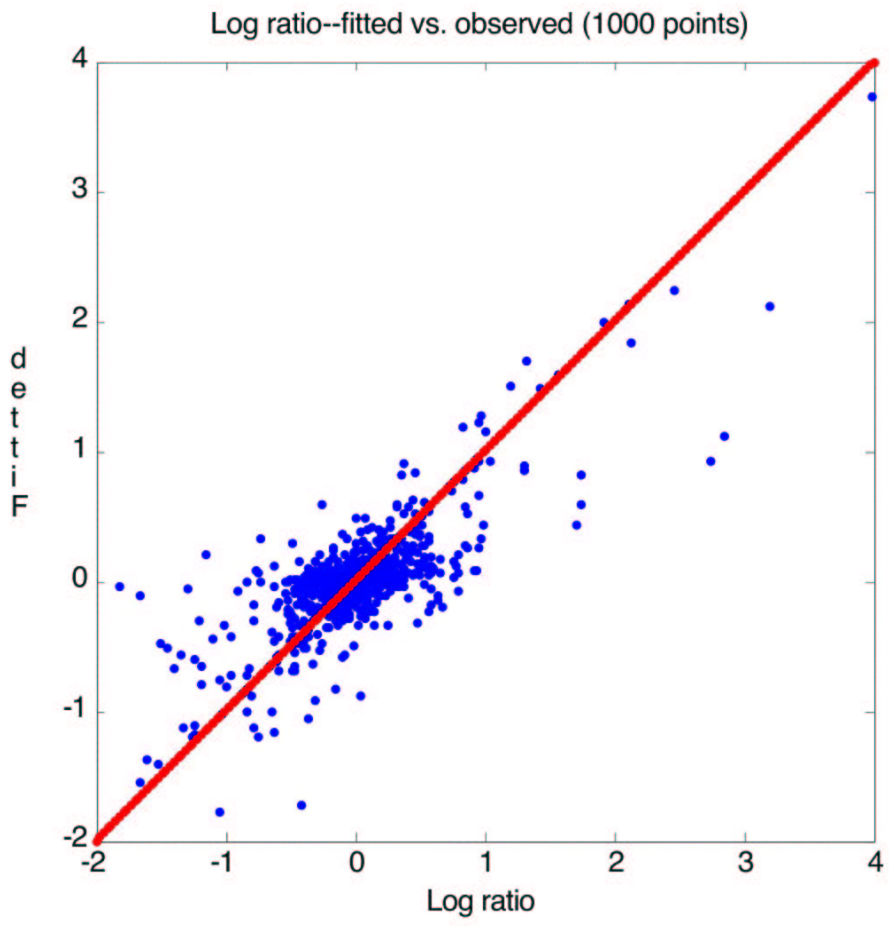
Rosetta Compendium
Causton stress
Cho mitotic
Chu sporulation
Epstein mitochondrial
Gasch Mec1
Gasch Stress
Hardwick Rapamycin
Lyons Zap1
Ogawa Phosphate
Spellman Cell Cycle
Zhu Forkhead

Results

Connection data taken from two separate sources: YPD and Young lab results

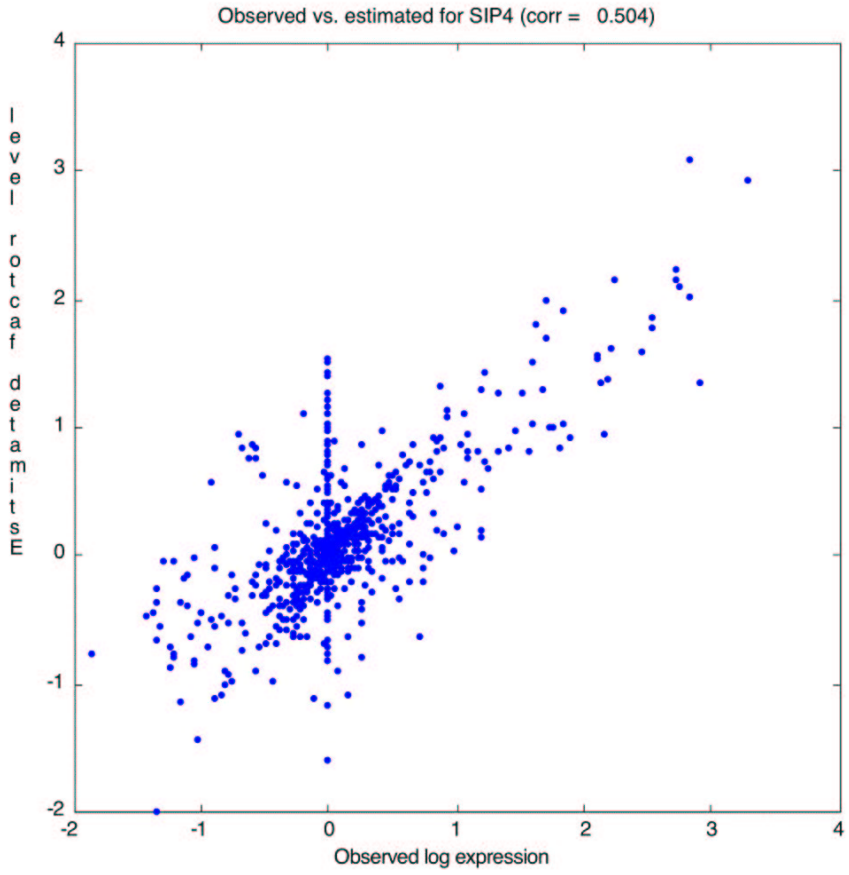
Reduction in overall variance:

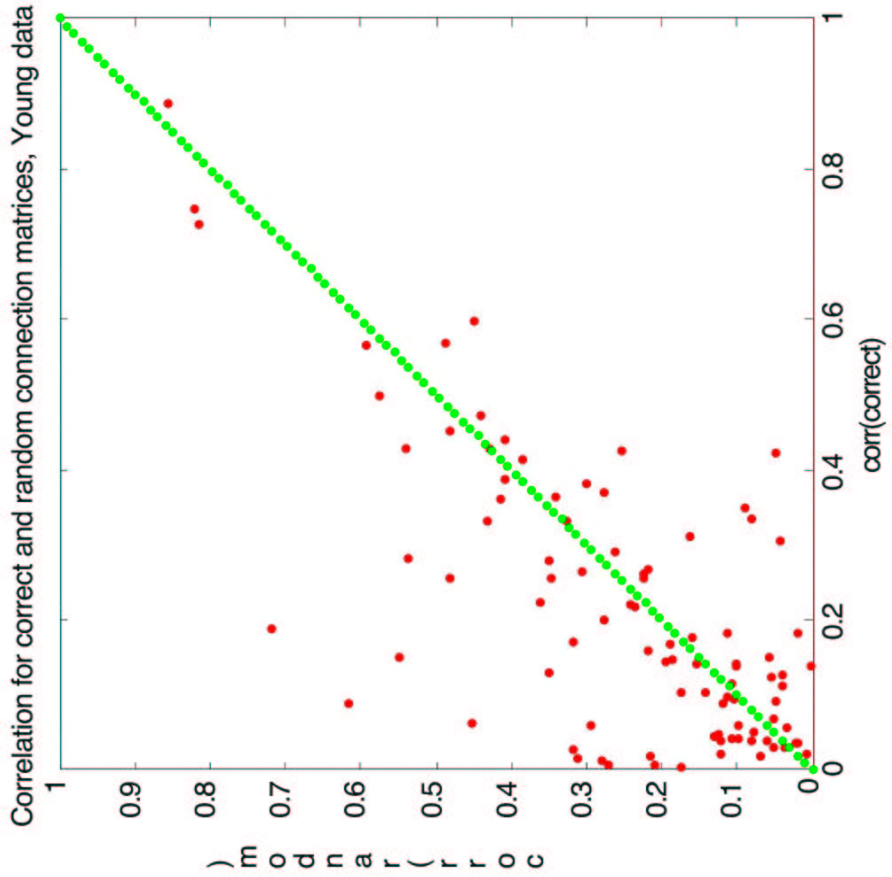
%Var expl(correct)%Var expl(random)

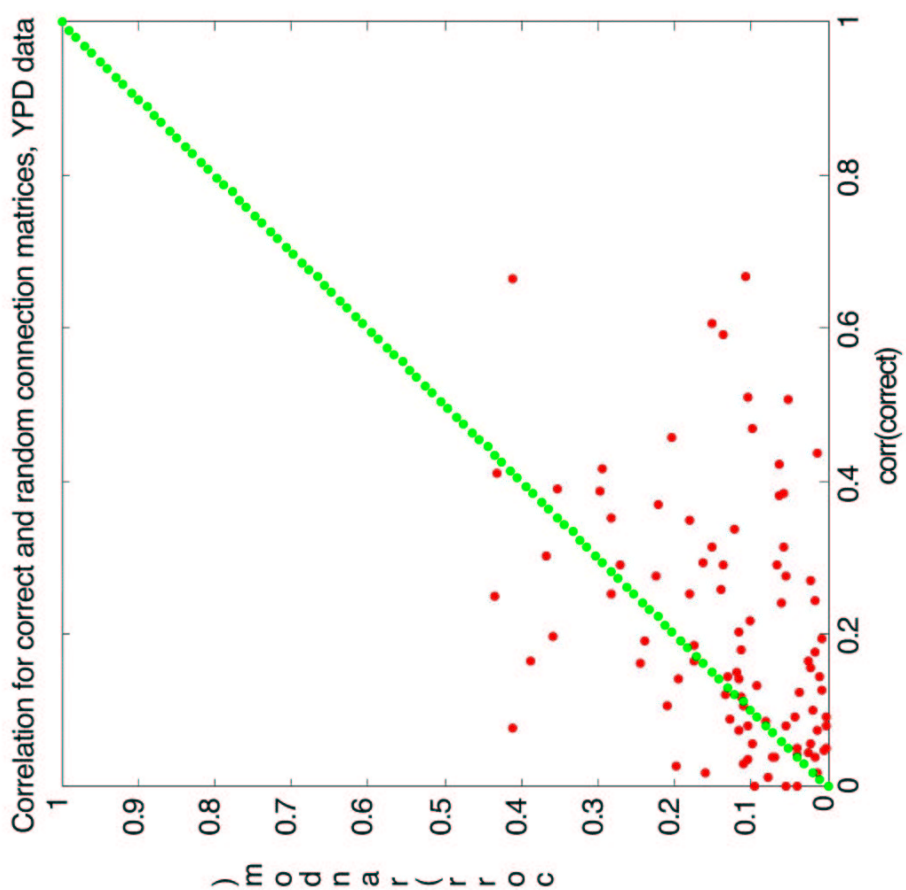


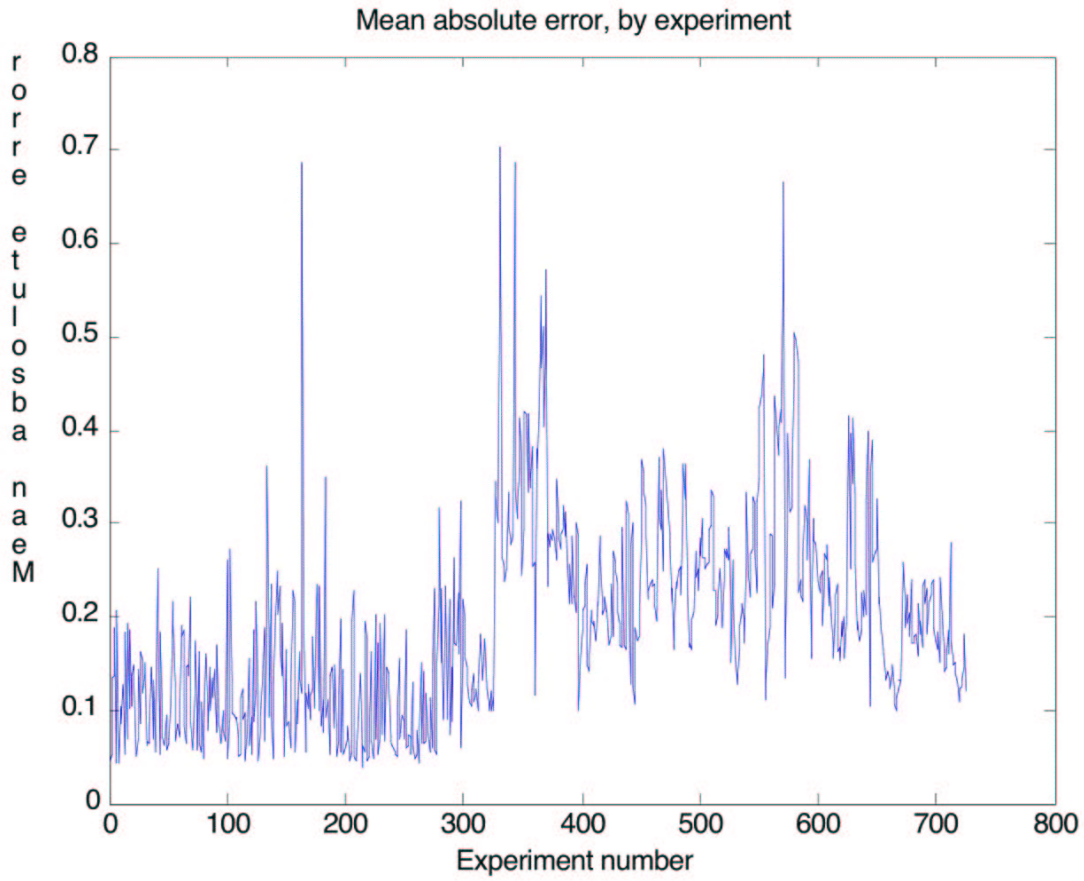
Results(2)--TF levels vs. transcript level

Studied correlation between the estimated level, X_f, of each transcription factor with the level of transcription of the corresponding gene, across all 725 experiments




Correlation for correct and random connection matrices, Young data
corr(correct)
corr(random)
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0.2
0.4
0.6
0.8


Correlation for correct and random connection matrices, YPD data
corr(correct)
corr(random)
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0.2
0.4
0.6
0.8



Conclusions/future work

We have demonstrated a flexible but simple model of transcription and shown that it can be solved rapidly and reproducibly.

The quality of the connection matrix is critical to the final results.

We are currently adding more high-quality connection information and additional expression data.

Future work (2)

We plan to study connections between "connection strength" (magnitude of α_{fg}) and other factors such as position, orientation, and binding site multiplicity.

We are actively seeking collaborators who have additional data or wish to try the model in other domains.