Dynamic models of regulatory networks

Michael R. Brent
Gene Regulatory Circuits

Transcription factor binding sites

- Computational
- Molecular: ChIP-seq, protein interaction, …

Johnson et al. 2007, *Science*
Gene Y

Transcription

RNA Transcript

Translation

Protein Y

DNA binding

Bound transcription factor Y

Gene X

Promoter

Promoter

Transcription start

Transcription start

(easy to measure)

(hard to measure)
Many systems for fully automated network reconstruction

Most use only RNA expression levels

- Microarrays have been around for a while
- Computer scientists love the uniformity

Most do not use

- Promoter sequence (TF factor binding motifs)
- *in vivo* TF binding data (ChIP/Seq)
- Protein-protein interaction data
Many systems for fully automated network reconstruction

Most output

• Yes/No: regulatory relationship between X & Y?
• Confidence level

Most do not output

• Direction of causality (who is the regulator?)
• Biochemical parameters
• Predictions for behavior in novel conditions, such as combinatorial knockout
The fundamental idea

- Measure RNA levels in many conditions (e.g. nutrients) and genetic perturbations (e.g. K/Os)
- Find correlations between genes
- If X is down whenever Y is up, perhaps Y represses X or X represses Y: X-Y

- Assume RNA=protein.
- No model of post-translational modifications.
Questions

Why are these tools so rarely used in practice (even by their developers!)?

• Do they really work? What can we expect?
• Which method, if any is useful?
• What expression data are most useful?
• Can better methods be developed?
Approach in three roles

**Benchmarkers**
- Objectively assess accuracy

**Algorithm developers**
- Improve the state of the art

**Users**
- Determine the unknown network that regulates induction of polysaccharide capsule in *C. neoformans* (fungal pathogen).
- Collaboration with Tamara Doering’s lab
Benchmarking with Artificial Gene Regulatory Networks

Generate a random network

Define its kinetics

Simulate its states

\[
\frac{\delta m_i}{\delta t} = S_i(R) - D_M m_i
\]

\[
\frac{\delta p_i}{\delta t} = T_P m_i - D_P p_i
\]

Add measurement noise

Infer network

Compare inferred to original
Assessing accuracy with A-BIOCHEM

Generate 50 random 100-gene networks
- Simulate steady-state expression levels in WT
- And 100 strains, each with one gene K/Oed.
- Assume accurate expression measurements

Compared three systems
- ARACNE
- CLR
- Symmetric-N
Evaluation Metric

Changing cutoffs for inclusion often results in a precision-recall tradeoff
Overall accuracy: A-BIOCHEM
How Realistic is the Benchmark?

A-BIOCHEM
• Distributions of number of targets per TF and TFs per target are off (degree)
• Biochemical parameters are arbitrary
• No translation/degradation: RNA=protein

GRENDEL (our benchmarking system)
• Degree distributions based on S. cerevisiae
• Biochemical parameters based on cerveisiae
• Translation/degradation model
Brian Haynes

Haynes & Brent (Bioinformatics, 2009)
Generating a realistic network

Degree distributions from *cerevisiae*

- Number of targets of a TF: \( \Pr(x) = x^{-0.69}, \ 0 < x < N \)
- Number of TFs regulating a target: \( \Pr(x) = e^{-0.39x}, \ 0 < x < N \)

Balaji et al. (2006)
Kinetic parameterization

HT experiments estimate, for each yeast gene,
- mRNA & protein degradation rates: $D^M, D^P$
- Translation rates: $T^P$
- Each simulated gene $i$ gets parameters of a randomly selected yeast gene (TFs use TF parameters)

Randomly generated network determines
- $R$: set of proteins regulating simulated gene $i$
- $S_i(R)$: Transcription rate of gene $i$ as a function of $R$

mRNA Rate Law

$$\frac{\delta m_i}{\delta t} = S_i(R) - D^M_i m_i$$

Protein Rate Law

$$\frac{\delta p_i}{\delta t} = T^P_i m_i - D^P_i p_i$$
Overall accuracy: GRENDEL

AUC=0.29

AUC=0.18
Effects of experimental design

Would more/better expression data help?

Same 50 hundred-gene networks

- Simulate steady-state expression levels in WT
  - 100 strains with each gene $\Delta$ in one strain
  - 100 strains with each gene overexpressed
  - Both data sets combined
Effect of Experimental Design

![Graph showing AUC-Precision Recall for ARACNE, CLR, and Symmetric-N with Knockouts, Overexpressed, and Knockouts + overexpressed categories.](image)
Inference from time courses

Should allow for better accuracy
  • Autoregulation not revealed by steady-state K/O

Simulated 33-hour time courses
  • Two external signals (e.g. glucose, Fe) perturbed twice at random times

Tested 2 systems that use such data
  • Symmetric-N, DBmcmc
Inference from time courses

Surprisingly little gain from increasing the sampling rate 30 fold with these two methods
Where do we stand?

Current systems may be useful for

- Hypothesizing a few, high-confidence regulatory relationships with a manageable number of incorrect hypotheses.
  - Hypotheses should be viewed skeptically
  - Experimental verification required

Current systems not useful for

- Reconstructing complete pathways
Improving the state of the art

Built a reconstruction system that

• generates testable predictions
  – distinguishes regulator from target
  – predicts behavior under novel conditions
• Can use steady-state and time course data
  – Models translation and degradation
• Is competitive in accuracy

Want a system that is more accurate!
Hope for improving accuracy

Think more like a molecular biologist

• Design experiments incrementally
• Classify genes by general function in advance
  – DNA binding, kinase, phosphatase, etc.
• Use more high-throughput data types
  – Promoter sequences (TF binding motifs)
  – \textit{In vivo} TF-to-DNA binding data (ChIP/seq)
  – Protein-protein interaction maps
Complementary approach

Build on molecular biology

• Let molecular biologists do what they do well
• Convert circles and arrows to quantitative, kinetic models
  – Goal is unchanged
• Write ODEs for all hypothesized reactions
• Measure kinetic parameters independently for each promoter, each TF, etc.
  – Enables predictions for rewired circuits
  – Important for synthetic biology/engineering
Quantifying and animating known Gene Regulatory Networks

Product of molecular biology

Write kinetic equations

\[
\frac{\delta m_i}{\delta t} = S_i(R) - D_M^i \cdot m_i
\]
\[
\frac{\delta p_i}{\delta t} = T_P^i \cdot m_i - D_P^i \cdot p_i
\]

Determine parameters

Test predictions

Predict static & kinetic behaviors
Method

Use approximate liberally
- A cell is a well mixed solution
- RNA and protein half-lives are constant
- No stochastic variation in reaction rates

Develop an algorithm/protocol
- Combine computational modeling with
- simple, repeatable, scalable experiments

Apply to multiple yeast pathways
- Test case: regulation of glucose transporters
Mark Johnston
Yeast ferment glucose to ethanol, storing solar energy and CO$_2$

Glucose $\rightarrow$ Yeast $\rightarrow$ Ethanol

$\text{CO}_2$ $\rightarrow$ Xylose $\rightarrow$ ??? $\rightarrow$ Glucose

$\text{CO}_2$ $\rightarrow$ Sun

$\text{CO}_2$ $\rightarrow$ Solar energy

$\text{CO}_2$ $\rightarrow$ Car emissions
Regulation of glucose transporters in *S. cerevisiae*
Parameters

- Vmax, 12
- Remaining, 14
- Glu uptake rates, 8
- transl. rates, 12
- protein deg. rates, 12
- mRNA deg. rates, 12

Optimized to steady state levels of HXT1,2,3 & 4 in wild type (in 2% Galactose and 2% Glucose).

From qPCR experiments done in our lab

Low throughput data

From high throughput data

\[
\frac{d[MTH1]}{dt} = V_{\text{max}} \left( \frac{1}{1 + \theta_1[Mig1]} \right) \left( \frac{1}{1 + \theta_2[Mig2]} \right) - D[MTH1]
\]

\[
0 = V_{\text{max}} \left( \frac{1}{1 + \theta_1[Mig1]} \right) \left( \frac{1}{1 + \theta_2[Mig2]} \right) - D[MTH1]
\]

\[
0 = V_{\text{max}} - D[MTH1]
\]
0 = V_{\text{max}} \left( \frac{1}{1 + \theta_1 [Mig1]} \right) \left( \frac{1}{1 + \theta_2 [Mig2]} \right) - D[MTH1]

0 = V_{\text{max}} \left( \frac{1}{1 + \theta_1 [Mig1]} \right) - D[MTH1]
Steady State Predictions

HXT1 levels

HXT2 levels

HXT3 levels

HXT4 levels
The graph illustrates the comparison of HXT4 WT over time (in minutes) with different glucose concentrations:

- **Red line**: 2% Gal to 2% Glu
- **Blue line**: 2% Gal to 0.1% Glu

The y-axis represents molecules/cell, and the x-axis represents time in minutes, ranging from 0 to 150.
Credits

My Lab
• Brian Haynes
• Sooraj Kuttykrishnan

Johnston Lab
• Jeff Sabina
Questions & Answers
\[ 0 = V_{\text{max}} \left( \frac{1}{1 + \theta_1[Mig1]} \right) \left( \frac{1}{1 + \theta_2[Mig2]} \right) - D[MTH1] \]