Dynamics on gene networks

Clayton W. Seitz

April 10, 2022

Outline

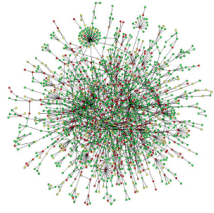
Computational network biology

Emerging research field that encompasses theory and applications of network models to study complex interactions of cells, DNA, RNA, proteins, and metabolites

Say we have a set of variables $\boldsymbol{x} = (x_1, x_2, ..., x_n)$ which might have some statistical dependence. $\boldsymbol{x}$ might be RNA or protein expression data, for example

- Often we are handed a batch of empirical samples $\boldsymbol{X} = \{\boldsymbol{x_1}, .., \boldsymbol{x_k}\}$
- We want to learn about the generating distribution $P(\boldsymbol{x}, t)$

Joint effort between physics, computer science, and biology



A gene interaction network

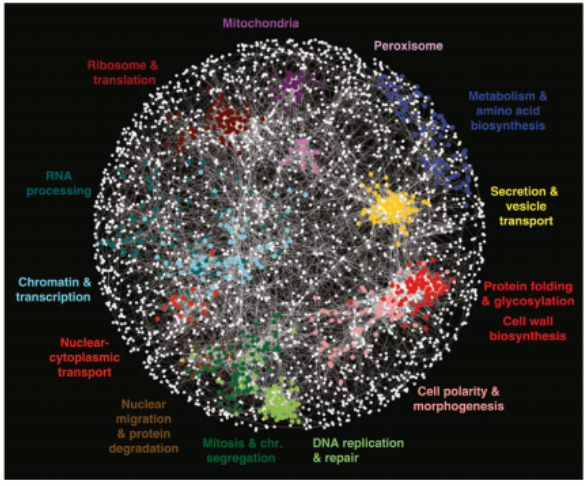
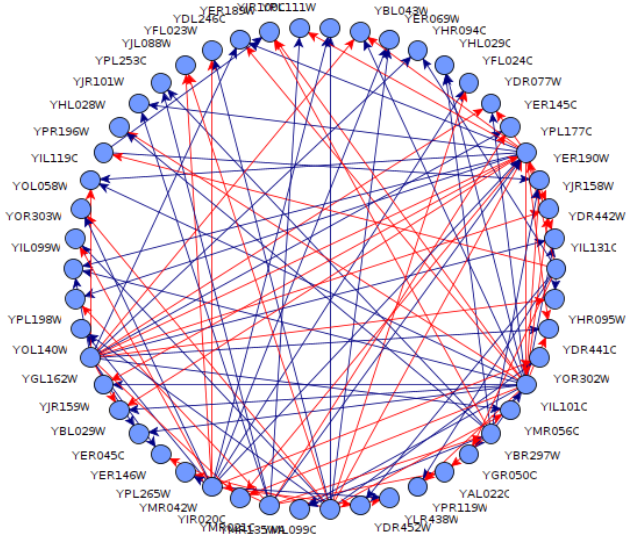


Figure 1: **Landscape of genetic interactions** in cells. Edges between genes denote Pearson correlation coefficients ($\rho > 0.2$) calculated from the complete genetic interaction matrix.

Example gene regulatory network in yeast



GNW
lis.epfl.ch/gnw

Experimental considerations

Gene interactions are inferred from gene expression data. RNA-seq has single cell-specificity and time resolution but lacks spatial resolution and data is noisy

FISH techniques have single-cell specificity, spatial resolution, less noisy, but multiplexing is difficult

High cost of multiplexing precludes acquisition of time-resolved data is single cell studies, which is important when statistics of the genes of interest are not stationary (circadian rhythms, cell-cycle, drug-treatment.

Even when we can collect single-cell time-series data, data collected at a time point will contain extra variability due to asynchrony of cells within a population (in terms of progression through a biological process)

Network modeling is hierarchical

Fine structure of molecular interactions sometimes can be resolved for low dimensionality

Computational complexity often scales exponentially with an increase in variables, density of interactions

In high-dimensional biological networks we often turn to classic dimensionality reduction or hidden variable models

Important topics

Models range from networks of a few genes with detailed dynamical models to very large networks with coarse statistical description.

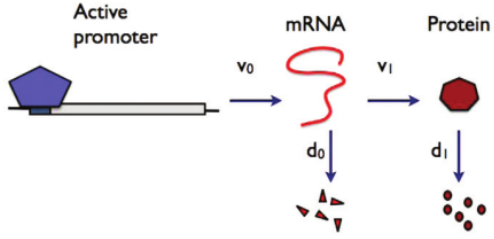
- Linear dynamics of small networks (deterministic, stochastic)
- Nonlinear dynamics of small networks (deterministic, stochastic) - Bintu model
- Inferring network structure - Phi-Mixing Coefficient
- Inferring network structure from linear dynamics - Hidden Markov Models
- Inferring nonlinear network structure from empirical data - ?
- Simulating stochastic dynamics (Gillespie algorithm)
- Simulating stochastic nonlinear dynamics (Michaelis-Menten kinetics, SERGIO)

Linear dynamics of transcription and translation

Assumptions: gene-gene interactions are linear, noise is Gaussian, long protein lifetimes

$$\dot{x}_i = \sum_j m_{ij} y_j - \alpha_i x_i + \eta_i$$

$$\dot{y}_i = r_i x_i - \beta_i y_i$$



If we assume that $\dot{y}_i \approx 0$ we have a Langevin equation for $x(t)$ and $y/x = \beta/r$

Let $\gamma_{ij} = m_{ij}\beta/r$. An example of a 3-gene system:

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_3 \end{bmatrix} = \begin{bmatrix} -\alpha_1 & \gamma_{21} & \gamma_{31} \\ \gamma_{12} & -\alpha_2 & \gamma_{32} \\ \gamma_{13} & \gamma_{23} & -\alpha_3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} + \begin{bmatrix} \eta_1 \\ \eta_2 \\ \eta_3 \end{bmatrix}$$

Ornstein-Uhlenbeck process

We have a linear SDE,

$$dx_i = \gamma_{ij} x_j dt + \sigma_{ij} dW$$

which has a corresponding Fokker-Planck equation:

$$\frac{\partial \tilde{P}(\vec{x}, t)}{\partial t} = -\boldsymbol{\gamma_{ij}} \frac{\partial}{\partial x_j} x_i \tilde{P}(\vec{x}, t) + D_{ij} \frac{\partial^2 \tilde{P}(\vec{x}, t)}{\partial x_i x_j} \tag{1}$$

If the real part of the eigenvalues of γ_{ij} are greater than zero, a stationary distribution exists

Conditional distributions of a Gaussian

Partition variables $\{x_n\}_{n=1}^N$ into sets $\boldsymbol{x}_a$ and $\boldsymbol{x}_b$.

$$\mu = \begin{bmatrix} \mu_a \\ \mu_b \end{bmatrix} \quad \Sigma = \begin{bmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{bmatrix}$$

The conditional distribution $p(\boldsymbol{x}_a|\boldsymbol{x}_b)$ must also be normal with parameters

$$\mu_{a|b} = \mu_a + \Sigma_{ab}\Sigma_{bb}^{-1}(x_b - \mu_b)$$
$$\Sigma_{a|b} = \Sigma_{aa} - \Sigma_{ab}\Sigma_{bb}^{-1}\Sigma_{ba}$$

Bayesian networks loosely express causal relationships. We can compare $p(\boldsymbol{x}_a|\boldsymbol{x}_b)$ and $p(\boldsymbol{x}_a)$. We can use this to assess quality of inference algorithms estimating the underlying network structure parameterized by the damping matrix Γ_{ij}

Marginal distributions of a Gaussian

The conditional distribution $p(x_1|x_2)$ between two variables $\boldsymbol{a} = x_1$, $\boldsymbol{b} = x_2$ has parameters

$$\mu_1 = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2)$$
$$\sigma_{1|2}^2 = \sigma_1^2 - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$$

The multivariate normal has the nice property that marginal distributions are

$$p(x_1) = \mathcal{N}\left(\mu_1, \sigma_1^2\right)$$

Conditional independence implies that $\mathcal{N}\left(\mu_1, \sigma_1^2\right) = \mathcal{N}\left(\mu_{1|2}, \sigma_{1|2}^2\right)$. We can then factor $p(\boldsymbol{x})$ into a Bayesian network.

A scalable algorithm for inferring gene regulatory networks

Interest in reverse-engineering whole-genome interaction networks from simultaneous measurements of the expression levels of all (or at least most) genes in many samples, under a common set of experimental conditions

This algorithm applies when it is valid to assume that the system is in a steady state. For systems out of equilibrium, we need inference algorithms designed to operate on time-series data

Typically gene expression data have low sampling rates and relatively small amount of data. Moreover, GRNs have a high number of genes with complex, nonlinear regulatory mechanisms

The Phi-Mixing Coefficient

Interest in reverse-engineering whole-genome interaction networks from simultaneous measurements of the expression levels of all (or at least most) genes in many samples, under a common set of experimental conditions

This algorithm applies when it is valid to assume that the system is in a steady state. For systems out of equilibrium, we need inference algorithms designed to operate on time-series data

Typically gene expression data have low sampling rates and relatively small amount of data. Moreover, GRNs have a high number of genes with complex, nonlinear regulatory mechanisms

References I