A brief introduction to graphical models and deep methods in computational network biology

Clayton W. Seitz

April 9, 2022

Outline

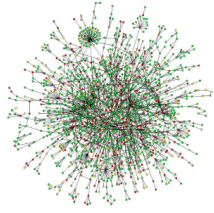
Computational network biology

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

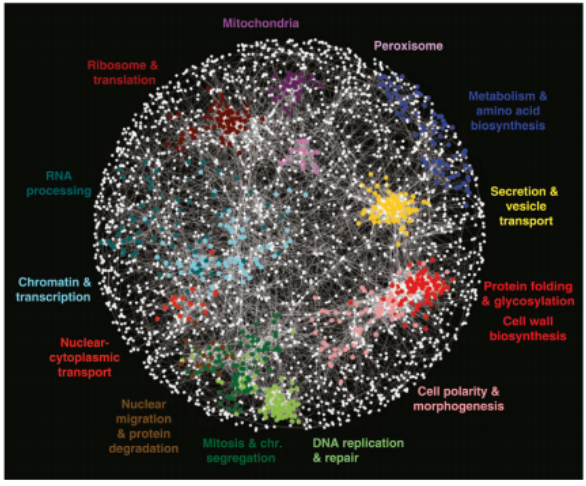


Figure 1: **Landscape of genetic interactions** in cells. Edges between genes denote Pearson correlation coefficients ($\rho > 0.2$) calculated from the complete genetic interaction matrix.

A protein interaction network

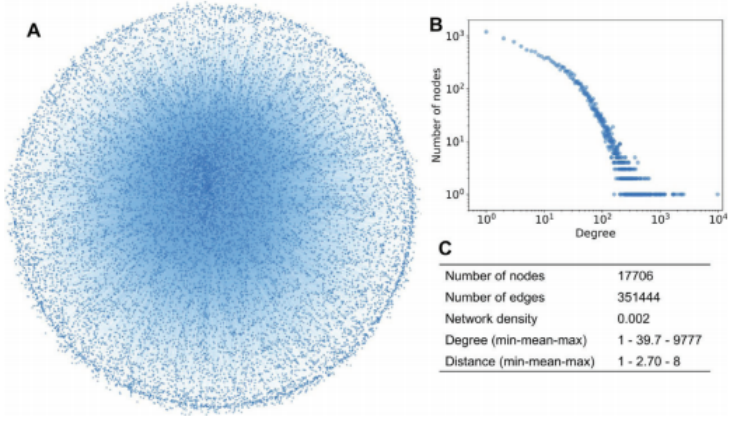


Figure 2: **Human protein interactome** of 17,706 proteins and 351,444 interactions (A) Overall complex network of human interactome. (B) Degree (connectivity) distribution of proteins by following a power-law tail. (C) Several selected network topological characteristics of the interactome.

A cellular interaction network (model neurons)

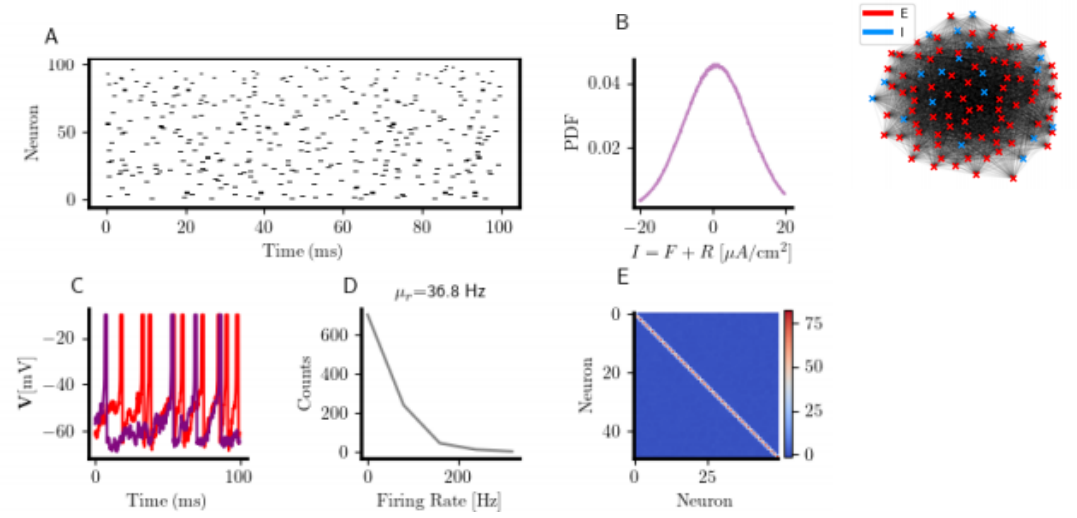


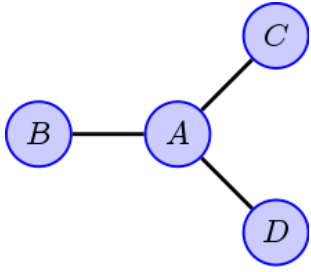
Figure 3: **Asychronous spiking of model neurons** (A) Steady-state raster plot of $N = 100$ uncoupled EIF neurons undergoing stimulation with GWN with $\mu = 2\mu A/\mathrm{cm}^2$ and $\sigma = 9\ \mu A/\mathrm{cm}^2$

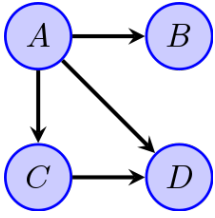
Probabilistic graphical models (PGMs)

Probabilistic graphical models are a class of machine learning algorithms that represent statistical dependencies of probability distributions as graphs

Two main types used in machine learning:
Bayesian Networks (BNs), Markov Random Fields (MRFs), but there are others

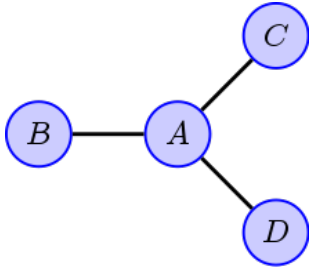
Major advantage is that they are structured models
They do not scale as easily as deep networks





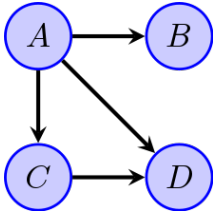
Probabilistic graphical models (PGMs)

Say we have a joint probability over gene expression $P(\boldsymbol{X})$
A PGM describes how $P(\boldsymbol{X})$ factors



Markov Random Fields (MRFs) e.g., Ising model

$$P(\boldsymbol{X};\Theta) = \frac{1}{Z}\prod_{i=1}^{N} P(\boldsymbol{X_i}, \mathcal{C}(X_i);\Theta_i)$$



Bayesian Network (BNs) - include causality

$$P(\boldsymbol{X}|\mathcal{G},\Theta) = \prod_{i=1}^{N} P(\boldsymbol{X_i}|\mathcal{C}(X_i),\Theta_i)$$

BNs as well as hybrid models have been used to examine gene expression

An example graphical model

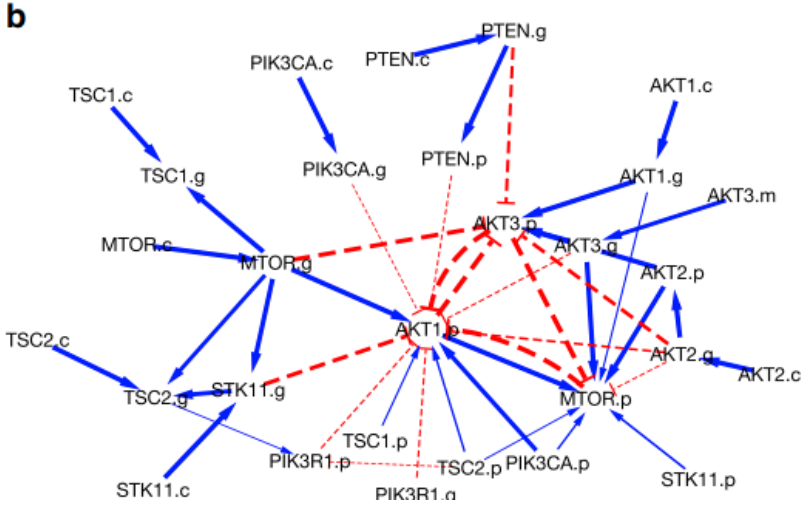


Figure 4: **PI3K pathway graph** discovery using graphical modeling (Ni et al. Bioinformatics 2018). c - transcript count, g - gene, p - protein, m - methylation

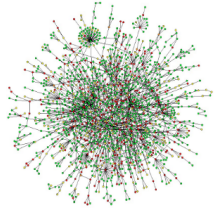
A tradeoff between mechanistic understanding and scale

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 deep methods

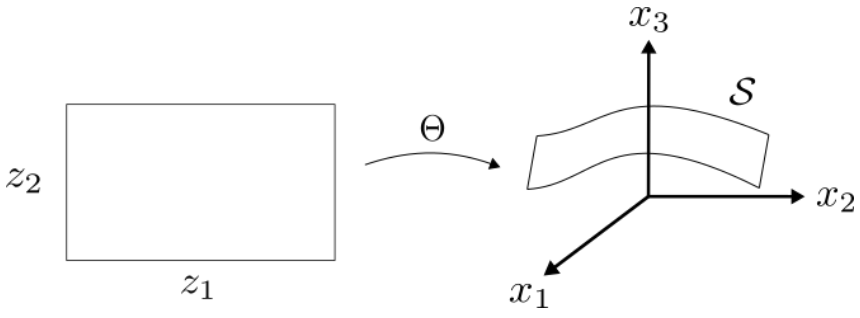
Introducing latent variables into the model can reduce computational complexity



Latent variables

Modeling all possible conditional dependencies quickly becomes intractable, lots of parameters

Introducing latent variables ***z*** can reduce the number of needed parameters



Variational autoencoders (VAEs)

The VAE architecture has been very succesful when applied to RNA-seq datasets see (Lopez Nature Methods 2020)

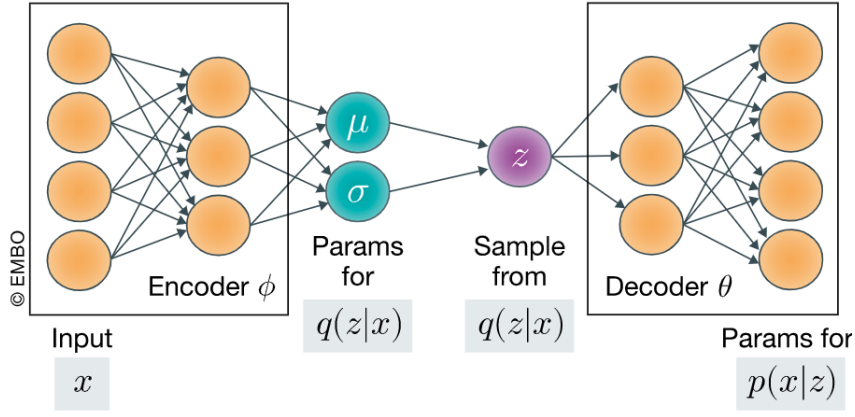
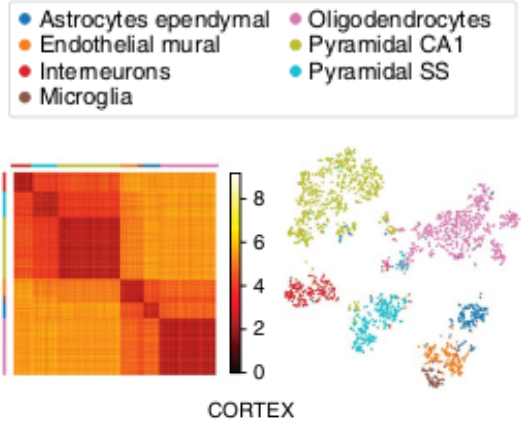


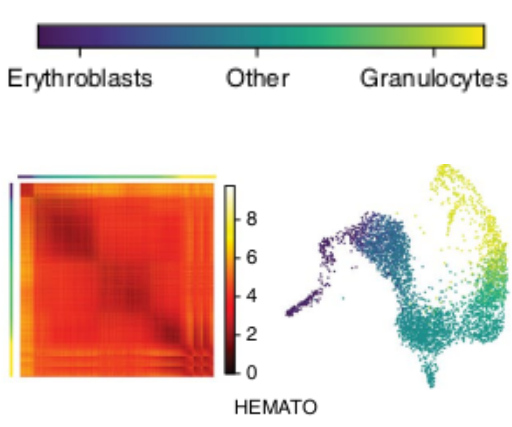
Figure 5: **Variational autoencoder architecture** Lopez 2020 EMBO

Using the VAE for cell phenotyping

Can do hypothesis testing (Bayes factor) but does not explicitly capture visible-visible causal relationships (captures latent-visible)

558 genes/3005 cells for CORTEX, 7,397 genes/4016 cells for HEMATO (Lopez Nature Methods 2020)





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 lack time resolution in single cell studies - RNA counts are not static (circadian rhythms, cell-cycle, drug-treatment)

in silico models make predictions based on kinetic parameters and gene regulatory networks, arbitrarily precise time resolution

Using *in silico* models to predict stationary statistics read out by multiplexed FISH? Or if dynamics are not relevant (maybe in tissue experiments), use graphical models on multiplexed FISH + spatial info

While we can collect single-cell time-series data, even data collected at one time point will contain variability due to (1) asynchrony of cells within a population (in terms of progression through a biological process), and (2) biological heterogeneity and often the presence of multiple cell (sub)types.

This means that the assumption that each cell is drawn iid is not necessarily valid

Partial information decomposition, variational autoencoder on static tissue data. Not confident that we can infer causality in this data format

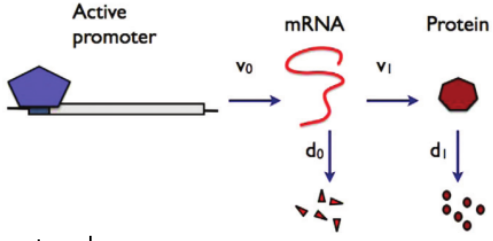
In-vitro, we can measure different time points which permits a larger class of models

Linear dynamics of transcription and translation

If we assume linearity (first-order reactions) in a gene regulatory network

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

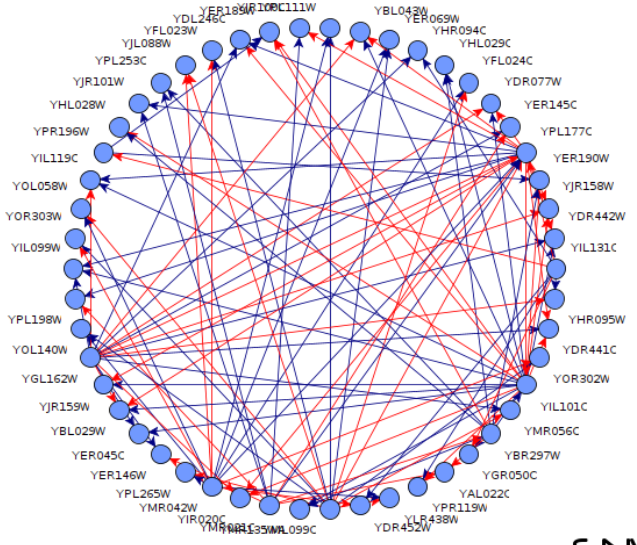
$$\dot{y}_i = r_i x_i - \beta_i y_i + \xi i$$



For example, a three-dimensional gene network:

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_3 \\ \dot{y}_1 \\ \dot{y}_2 \\ \dot{y}_3 \end{bmatrix} = \begin{bmatrix} -\alpha_1 & 0 & 0 & m_{11} & m_{21} & m_{31} \\ 0 & -\alpha_2 & 0 & m_{12} & m_{22} & m_{32} \\ 0 & 0 & -\alpha_3 & m_{13} & m_{23} & m_{33} \\ r_1 & 0 & 0 & -\beta_1 & 0 & 0 \\ 0 & r_2 & 0 & 0 & -\beta_2 & 0 \\ 0 & 0 & r_3 & 0 & 0 & -\beta_3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ y_1 \\ y_2 \\ y_3 \end{bmatrix} + \begin{bmatrix} \eta_1 \\ \eta_2 \\ \eta_3 \\ \xi_1 \\ \xi_2 \\ \xi_3 \end{bmatrix}$$

Example gene regulatory network in yeast



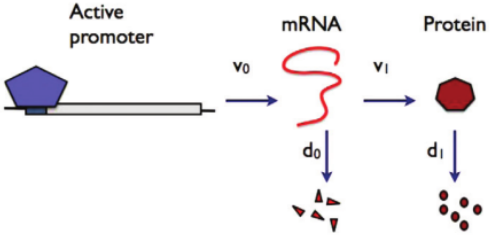
GNW
lis.epfl.ch/gnw

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.

Handling systems out of equilibrium

In practice we cannot necessarily assume that the data is Gaussian or that the distribution is stationary. It is very difficult to guarantee equilibrium as we can for simulations

We also may need to perform preprocessing on the data to determine if there is population heterogeneity (VAE?)

References I