Statistical Issues in Reverse Engineering of Genetic Networks
Poster at PSB 98
John Hertz
Nordita, Copenhagen

background
How many knockouts do we have to do to determine the genome? More generally, how many expression trajectories, of what lengths, etc, do we need in order to find with what strength all genes are regulated by all others, or, at least, in order to be able to predict the outcome of any bit of genetic engineering? This is a complex statistical problem. Here I try to explore the basic issues in a simple model.

the model
Suppose we have \( N \) genes in a regulatory network, and we know that on average each one is regulated by \( K \) others but not which ones. We have data from experiments, each of which tells us that in some situation a particular expression pattern is followed by another one. To make the problem tractable, let us assume that the regulation is always additive, that the expression state of every gene can be classified as either “on” \( (S_i = 1) \) or “off” \( (S_i = -1) \), and that we can unambiguously identify a suitable “time unit” over which the change in expression pattern occurs. The network is then governed by
\[
S_i(t + 1) = \text{sgn} \left[ \sum_j J_{ij} S_j(t) \right].
\] (1)
The problem in genomic biology is: given a set of initial and final states \( S_i(t), S_i(t + 1) \), find the \( J_{ij} \)'s. I will not solve this problem. Instead, I will address the problem: how many such data are necessary (on average) to find the \( J_{ij} \)'s well enough to predict, with some specified level of confidence, the next step from a new initial state.

In this problem, one can look at the outcome one gene at a time (a single value of the index \( i \) in (1), i.e., we are talking about the influences of all other genes on this one. Furthermore, it is a fact that the genome is not densely
connected – most of the $J_{ij}$’s are zero. We just do not know which particular ones are not. So we write (dropping the $i$ index)

$$J_j = n_j w_j,$$

where the $n_j$ are 0 or 1, subject to the constraint

$$\sum_j n_j = K.$$  

We will eventually use the fact that $K \ll N$.

Since we have $O(N)$ unknown parameters, the fear is that it will take $O(N)$ experiments to determine them. The analysis to be given here suggests that the situation is not quite that bad.

the formal problem

Suppose the genome that generated the data had connection weights $w_j = w_j^0$, and call its $n_j n_j^0$. In learning theory, we call this the “teacher” network and the one whose parameters we are trying to fit the “student”. The rules of the game are that we do not know the teacher’s parameters, only some examples of the changes in expression they lead to from specific initial conditions. The question to be answered is: of all the possible gene networks specified by sets of $w_j$ and $n_j$, what fraction of them make the same prediction for the next step for our whole collection of $p$ different initial states (which we denote by $\xi^\mu_j$, $\mu = 1, \cdots, p$). This is a normalized sum of the volumes in $w$ space where the student and teacher agree:

$$V(w^0, n^0, \xi) = \frac{1}{\mathcal{N}} \sum_{\{n_j\}} \int \prod_i \frac{dw_i}{\sqrt{2\pi}} \exp\left[-\frac{1}{2} \sum_j w_j^2\right] \delta_{\sum_i n_i, K}$$

$$\times \prod_{\mu = 1}^p \Theta(K^{-1/2} w \cdot n \cdot \xi^\mu - K^{-1/2} w^0 \cdot n^0 \cdot \xi^\mu),$$

where $a \cdot b \cdot c$ means $\sum_j a_j b_j c_j$. (Instead of putting in delta functions to enforce constraints on the weight magnitudes, I have just put a Gaussian weight distribution with unit variance. I have checked that putting in the delta functions explicitly gives the same results.) The normalizing denominator

$$\mathcal{N} = \sum_{\{n_j\}} \delta_{\sum_i n_i, K}$$

$$2$$
cancels out in the end, so I won’t say anything else about it.

The way to think about this complicated-looking integral is the following. Every added input-output example tends to restrict the volume 4) further. For an infinite number of examples, the volume reduces to the point $w = w^0$, $n = n^0$. For large but finite $p$ there will typically be a small volume in $w$, $n$ space where the output of the student network, with connections specified by $w$ and $n$, agrees with that of the teacher network with connections given by $w^0$ and $n^0$.

A measure of how closely the student network resembles the teacher is

$$R = \frac{1}{K} \sum_j \langle w_j n_j w^0_0 n^0_0 \rangle,$$

where the average is over the allowed volume $V(w^0, n^0, \tilde{\xi})$ in (5) and over all choices of the example inputs $\xi_j^\mu$. In this analysis I assume these are chosen completely randomly and independently in the state space of the network. Real data are likely to be correlated, and a future treatment should address this problem, but for now I just do the simple analysis based on independent random examples.

The volume $V(w^0, n^0, \tilde{\xi})$ fluctuates wildly from one set of examples to another. Its log has much weaker fluctuations, so to get typical answers we average that over the $\xi_j^\mu$. For convenience we will also average over all teachers consistent with the general constraint (3). Using the identity $\log x = \lim_{m \to 0} d(x^m)/dx$, we get

$$\langle \log V \rangle = \lim_{m \to 0} \frac{\partial \langle V^m \rangle}{\partial m}$$

with

$$\langle V^m \rangle = \frac{1}{N^m} \sum_{\{n_j\}} \int D\delta \sum_{j} n_j, K \prod_{\mu=1}^{p} \prod_{\alpha=1}^{m} \Theta(K^{-1/2} w^\alpha \cdot n^\alpha \cdot \tilde{\xi}^\mu \cdot K^{-1/2} w^0_0 n^0_0 \cdot \tilde{\xi}^\mu),$$

where

$$\int D w \equiv \int \prod_{j, \alpha} D w_j^\alpha = \int \left[ \prod_{j, \alpha} \frac{dw_j^\alpha e^{-(w_j^\alpha)^2/2}}{\sqrt{2\pi}} \right].$$

\footnote{This is a standard method in the statistical mechanics of disordered systems. It is called the “replica method” because it corresponds formally to taking $m$ copies (replicas) of the system.}
evaluating it

The manipulations are standard in neural network theory; they go back to the classic work of Elizabeth Gardner[1] and are only a modest extension of the treatment by Györgyi and Tishby[2] of the undiluted version of this problem. I summarize them here for the benefit of readers not familiar with similar problems who want to get an idea of how the calculation is carried out. Readers not interested in the formal details, can skip forward to the analysis following Eqns. (42) and (43).

In the numerator, we have $p$ independent averages, one for each pattern, equal to:

$$A = \left\langle \prod_{\alpha=1}^{m} \Theta(K^{-1/2}w^\alpha \cdot n^\alpha \cdot \xi^\alpha \cdot K^{-1/2}w^0 \cdot n^0 \cdot \xi^0) \right\rangle =$$

$$\left\langle \left[ \prod_{\alpha=0}^{m} \Theta(K^{-1/2}w^\alpha \cdot n^\alpha \cdot \xi^\alpha) + \prod_{\alpha=0}^{m} \Theta(-K^{-1/2}w^\alpha \cdot n^\alpha \cdot \xi^\alpha) \right] \right\rangle,$$  \hspace{0.5cm} (10)

where we have used $\Theta(ab) = \Theta(a)\Theta(b) + \Theta(-a)\Theta(-b)$, $\Theta^2(x) = \Theta(x)$, and $\Theta(x)\Theta(-x) = 0$. Notice that now everywhere the teacher (index $\alpha = 0$) appears on the same footing as the students everywhere in the problem, i.e., all integrals over $w_j^\alpha$'s are over all $m + 1$ of them, and similarly for the sums over the $n_j^\alpha$'s. (In the otherwise extensive treatment of this problem by Kuhlmann and Müller [3], this symmetric case was not studied.)

Now use the integral representation of the $\Theta$-functions,

$$\Theta(z) = \int_{0}^{\infty} d\lambda \delta(\lambda - z) = \int_{0}^{\infty} d\lambda \int_{-\infty}^{+\infty} \frac{dx}{2\pi} e^{i\lambda(x-z)}$$  \hspace{0.5cm} (11)

to write (10) as

$$A = 2 \int \prod_{\alpha} \frac{d\lambda_{\alpha} dx_{\alpha}}{2\pi} \exp(i \sum_{\alpha} x_{\alpha} \lambda_{\alpha}) \langle \cos(K^{-1/2} \sum_{\alpha} x_{\alpha} \sum_{j} w_j^\alpha n_j^\alpha \xi_j) \rangle.$$  \hspace{0.5cm} (12)

(From now on, all $d\lambda_{\alpha}$ integrals are meant to run from 0 to $\infty$.) For $K \gg 1$ we can expand the cosine to quadratic order, reexponentiate, and perform the average over $\xi_j = \pm 1$ to give

$$A = 2 \int \prod_{\alpha} \frac{d\lambda_{\alpha} dx_{\alpha}}{2\pi} \exp(i \sum_{\alpha} x_{\alpha} \lambda_{\alpha}) \exp \left( \frac{1}{2K} \sum_{\alpha \beta} x_{\alpha} x_{\beta} \sum_{j} w_j^\alpha n_j^\alpha w_j^\beta n_j^\beta \right).$$  \hspace{0.5cm} (13)
We now define the order parameter matrix \( q_{\alpha \beta} \) as
\[
q_{\alpha \beta} = \frac{1}{K} \sum_{j} w_j^\alpha n_j^\alpha w_j^\beta n_j^\beta.
\] (14)

Note that the average of \( q_{0\alpha} \) is the the quantity \( R \) introduced in (6).

Eq. (13) can be written
\[
A = 2 \int \prod_{\alpha} \frac{d\lambda_{\alpha} dx_{\alpha}}{2\pi} \exp \left( i \sum_{\alpha} x_{\alpha} \lambda_{\alpha} \right) \exp \left( \frac{1}{2} \sum_{\alpha \beta} x_{\alpha} x_{\beta} q_{\alpha \beta} \right).
\] (15)

We can evaluate this in exactly the same way as in the undiluted problem. Assuming replica symmetry (\( q_{\alpha \beta} = q \) for all \( \alpha \neq \beta, q_{\alpha \alpha} = 1 \)), we have
\[
\frac{1}{2} \sum_{\alpha \beta} q_{\alpha \beta} x_{\alpha} x_{\beta} = \frac{1}{2}q \left( \sum_{\alpha} x_{\alpha} \right)^2 + \frac{1}{2}(1 - q) \sum_{\alpha} x_{\alpha}^2,
\] (16)

and using the Gaussian integral identity
\[
\exp \left( -\frac{1}{2}a^2 \right) = \int \frac{dt}{\sqrt{2\pi}} \exp \left( -\frac{1}{2}t^2 + iat \right) = \int \text{D}t \, e^{iat}
\] (17)

we get
\[
A = 2 \int \text{D}t \int \prod_{\alpha} \frac{d\lambda_{\alpha} dx_{\alpha}}{2\pi} \exp \left[ i \sum_{\alpha} \left(x_{\alpha} (\lambda_{\alpha} + t\sqrt{q})\right) \right] \exp \left[ -\frac{1}{2} (1 - q) \sum_{\alpha} x_{\alpha}^2 \right]
\]
\[
= 2 \int \text{D}t \, I^{m+1}(q, t),
\] (18)

with
\[
I(q, t) = \int \frac{dx d\lambda}{2\pi} \exp \left[ i x(\lambda + t\sqrt{q}) - \frac{1}{2} (1 - q) x^2 \right].
\] (19)

This integral can be put into closed form, completing the square in the exponent, performing the \( dx \) integral, and making a linear shift of variable in the remaining \( d\lambda \) integral to give
\[
I(q, t) = \text{H} \left( t \frac{\sqrt{q}}{1 - q} \right),
\] (20)

where
\[
\text{H}(x) = \int_{x}^{\infty} \text{D}x = \int_{x}^{\infty} \frac{dx}{\sqrt{2\pi}} \exp \left( -\frac{1}{2} x^2 \right).
\] (21)
Recalling that there are $p$ initial expression patterns to average over, we have a net factor of

$$A^p = \left[ 2 \int D t H^{m+1} \left( t \sqrt{\frac{q}{1-q}} \right) \right]^p. \quad (22)$$

In the $m \to 0$ limit, we can write

$$A = 2 \int D t H \left( t \sqrt{\frac{q}{1-q}} \right) \left[ 1 + m \ln H \left( t \sqrt{\frac{q}{1-q}} \right) \right], \quad (23)$$

so, using the fact that $H(x)$ is $\frac{x}{2}$ plus an odd function of $x$, we get

$$A = 1 + 2m \int D t H \left( t \sqrt{\frac{q}{1-q}} \right) \ln H \left( t \sqrt{\frac{q}{1-q}} \right), \quad (24)$$

or

$$\ln A^p = 2mp \int D t H \left( t \sqrt{\frac{q}{1-q}} \right) \ln H \left( t \sqrt{\frac{q}{1-q}} \right). \quad (25)$$

We have now performed the average over the $\xi_j$ in the numerator, and we see that the result depends on the $w_j^\alpha$ and $n_j^\alpha$ only through the quantity $q_{\alpha\beta}$. The next step is to convert the integrals over $w_j^\alpha$ and sums over $n_j^\alpha$ to one over $q_{\alpha\beta}$. We do this by means of delta functions:

$$1 = \int \prod_{\alpha<\beta} dq_{\alpha\beta} K \delta(Kq_{\alpha\beta} - \sum_j w_j^\alpha n_j^\alpha w_j^\beta n_j^\beta)$$

$$= \int \prod_{\alpha<\beta} \frac{dq_{\alpha\beta}}{2\pi i} F_{\alpha\beta} K \exp[-F_{\alpha\beta}(Kq_{\alpha\beta} - \sum_j w_j^\alpha n_j^\alpha w_j^\beta n_j^\beta)]. \quad (26)$$

The other constraint (3) is imposed by the factor

$$\prod_{\alpha} \delta \sum_j n_j^\alpha, K = \prod_{\alpha} \int_{-i\pi}^{i\pi} \frac{d\mu_\alpha}{2\pi i} \exp[\mu_\alpha(K - \sum_j n_j^\alpha)]. \quad (27)$$

The fact that all the $w_j^\alpha$'s appear quadratically in the exponent (and that different $j$ do not mix) now allows all the $dw_j^\alpha$ integrals to be performed. For each $j$ we have to evaluate

$$J(F, \mu) = \sum_{\{n_j^\alpha\}} \int \prod_{\alpha} \frac{dw_j^\alpha}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2} \sum_\alpha (w_j^\alpha)^2 + \frac{1}{2} \sum_{\alpha \neq \beta} F_{\alpha\beta} w_j^\alpha n_j^\alpha w_j^\beta n_j^\beta - \mu_\alpha n_j^\alpha \right\}. \quad (28)$$
Again assuming replica symmetry ($\mu_\alpha = \mu$ and $F_{\alpha\beta} = F$) and applying the Gaussian integral identity we get

$$J(F, \mu) = \int \text{D}t \left[ 1 + e^{-\mu} J_2(F, t) \right]^{m+1}, \quad \text{(29)}$$

where

$$J_2(F, t) = \int \frac{dw}{\sqrt{2\pi}} \exp\left[ t \sqrt{F} w - \frac{1}{2}(1 + F)w^2 \right] = \frac{1}{\sqrt{1 + F}} \exp \left[ \frac{t^2 F}{2(1 + F)} \right]. \quad \text{(30)}$$

Expanding for small $m$,

$$J(F, \mu) = \int \text{D}t \left[ 1 + e^{-\mu} J_2(F, t) \right] \left[ 1 + m \ln(1 + e^{-\mu} J_2(F, t)) \right]$$

$$= 1 + e^{-\mu} + m \int \text{D}t \left[ 1 + e^{-\mu} J_2(F, t) \right] \ln[1 + e^{-\mu} J_2(F, t)]$$

$$= (1 + e^{-\mu}) \left[ 1 + \frac{m}{1 + e^{-\mu}} \int \text{D}t \left[ 1 + e^{-\mu} J_2(F, t) \right] \ln[1 + e^{-\mu} J_2(F, t)] \right]. \quad \text{(31)}$$

Thus

$$\ln J(F, \mu) = \ln(1 + e^{-\mu})$$

$$+ \frac{m}{1 + e^{-\mu}} \int \text{D}t \left[ 1 + e^{-\mu} J_2(F, t) \right] \ln[1 + e^{-\mu} J_2(F, t)]. \quad \text{(32)}$$

We now have an integral over the $q_{\alpha\beta}$, $F_{\alpha\beta}$, and $\mu_\alpha$ of $A^p$ (from (22)), $J^N$ (with $J$ from (29)), and the exp($-KF_{\alpha\beta}q_{\alpha\beta}$) and exp $K\mu_\alpha$ factors in (26), and (27), respectively. Calling the integrand exp $\Omega(q, F, \mu)$, we have

$$\Omega(q, F, \mu) = 2mp \int \text{D}t H(tQ) \ln H(tQ)$$

$$+ N \ln(1 + e^{-\mu}) + \frac{Nm}{1 + e^{-\mu}} \int \text{D}t \left[ 1 + e^{-\mu} J_2(F, t) \right] \ln[1 + e^{-\mu} J_2(F, t)]$$

$$- \frac{1}{2} Km(m + 1)Fq + K(m + 1)\mu, \quad \text{(33)}$$

with

$$Q = \sqrt{\frac{q}{1 - q}}. \quad \text{(34)}$$

All the terms in $\Omega$ are proportional to large numbers ($p$, $K$, or $N$), so the integrals are dominated by the stationary points with respect to $q$, $F$, and
\( \mu^2 \) The quantity we are most interested in is the saddle point value of \( q \), because it is equal to \( R \) (Eqn. (6)), which tells us how accurately the true weights are determined by the available data.

First we find the stationary value of the chemical potential \( \mu \). It is sufficient to work to lowest order in \( e^{-\mu} \). Expanding \( \log J(F, \mu) \) in \( e^{-\mu} \) gives

\[
\ln J(F, \mu) = \ln(1 + e^{-\mu}) + \frac{m}{1 + e^{-\mu}} \int dt [1 + e^{-\mu} J_2(F, t)] [e^{-\mu} J_2(F, t) - \frac{1}{2} e^{-2\mu} J_2^2(F, t)] = (1 + m)e^{-\mu} + \mathcal{O}(e^{-2\mu}).
\]

Thus the \( \mu \)-dependent part of \( \Omega \) is, to this order,

\[ \Omega_0(\mu) = N(m + 1)e^{-\mu} + K(m + 1)\mu, \]

and stationarity with respect to \( \mu \) gives

\[ e^{-\mu} = K/N \approx f. \]

We now look at the rest of \( \Omega \), which is linear in \( m \) in the limit \( m \to 0 \). Calling it \( K\mu G(q, E, F) \), we have

\[
G(q, E, F) = 2\alpha \int dt H(tQ) \ln H(tQ) + \frac{1}{f} \int dt [1 + e^{-\mu} J_2(F, t)] \ln [1 + e^{-\mu} J_2(F, t)] - \frac{1}{2} F q,
\]

with

\[ \alpha = p/K, \]

the number of data relative to the connectivity. The saddle point equations are

\[ \frac{1}{2} F = 2\alpha \frac{\partial}{\partial q} \int dt H(tQ) \ln H(tQ) \]

and

\[ \frac{1}{2} q = \frac{1}{f} \frac{\partial}{\partial F} \int dt [1 + e^{-\mu} J_2(F, t)] \ln [1 + e^{-\mu} J_2(F, t)] \]

\(^2\)Normally we would look for maxima of \( \Omega \). However, here, because the dimensionality of the \( q \)-space is negative, it turns out that we want minima.
After some algebra they can be put into the forms

\[
F = \frac{\alpha}{\pi(1 - q)^{1/2}} \int \frac{dt}{\Delta t} \frac{\exp(-\frac{1}{2}q t^2)}{H(t, \sqrt{q})}
\]  
(42)

and

\[
q = \frac{1}{(1 + F)} \int \frac{du}{\sqrt{2\pi}} e^{-u^2/2} [u^2 - 1] \log \left( 1 + \frac{f}{\sqrt{1 + F} e^{u^2F/2}} \right).
\]  
(43)

**solution: limiting cases**

We deal first with (43). For large \( F \), we can ignore the 1 term inside the log, leading to

\[
q = \frac{1}{1 + F} \int Du(u^2 - 1)[-\mu - \frac{1}{2} \log(1 + F) + \frac{1}{2}u^2F] = \frac{F}{1 + F}
\]  
(44)

This is the same equation one gets in the undiluted case. For small \( F \), we can expand the log and find

\[
q = \frac{1}{(1 + F)^{3/2}} \int \frac{du}{\sqrt{2\pi}} e^{-u^2/2} [u^2 - 1] \frac{f}{\sqrt{1 + F} e^{u^2F/2}}
\]

\[
= \frac{f}{(1 + F)^{3/2}} \int \frac{du}{\sqrt{2\pi}} e^{-u^2(1-F)/2} (u^2 - 1)
\]

\[
= \frac{f F}{(1 - F^2)^{3/2}} \approx f F.
\]  
(45)

The right hand side of (43) can be expanded around these limits in powers of \( f/F \) or \( F/f \), respectively. Numerical integration using Mathematica shows a smooth transition between the two regimes.

Now have a look at the other saddle point equation, (42). At small \( q \), the integral on the right hand side approaches the value 2. It is a smooth function of \( q \); a linear expansion in \( q \) is quite accurate all the way up to \( q = 1 \): 

\[
g(q) \equiv \int \frac{dt}{\Delta t} \frac{\exp(-\frac{1}{2}q t^2)}{H(t, \sqrt{q})} \approx 2 + \left( \frac{4}{\pi} - 1 \right) q;
\]  
(46)

\[
g(1) \approx 2.273,
\]  
(47)
so the right hand side of (42) starts out at the value $\alpha/\pi$ for small $q$ and rises with $q$, diverging like $0.72\alpha/\sqrt{1-q}$ as $q \to 1$.

There are thus two interesting limits: small and large $\alpha$. For small $\alpha$, the curve $F = \alpha g(q)/(\pi \sqrt{1-q})$ intersects $q = f F$ at

$$q = \frac{2\alpha f}{\pi}. \quad (48)$$

This is exactly the same result as in the undiluted problem if we had $\alpha = p/N$ instead of $\alpha = p/K$. That is, at very small $\alpha$, the behaviour of $q(\alpha)$ just reflects the fact that we have $N$ weights to find. In the opposite limit, the equations are exactly the same as in the undiluted problem, so the asymptotic approach of $q(\alpha)$ to 1 is as if we knew in advance which weights were zero. In between there is a smooth crossover. The behaviour deviates significantly from the small-$\alpha$ limiting form already below $\alpha = 1$, but does not reach the large-$\alpha$ limiting form until $\alpha f \gg 1$. The form of (43) suggest that the crossover occurs for $\alpha = O[\log(1/f^2)]$, but it is in fact very broad.

Another measure of how well the true network parameters are determined by the examples is the generalization error $\epsilon_g$, defined as the probability of giving the wrong output on an example not used in fitting the $w_j$ and $n_j$. Elementary geometrical arguments relate $\epsilon_g$ to $q$:

$$\epsilon_g = \frac{1}{\pi} \cos^{-1} q. \quad (49)$$

In our two limits we get

$$\epsilon_g \approx \frac{1}{2} - q/\pi = \frac{1}{2} \left( 1 - \frac{2\alpha f}{\pi^2} \right) \quad (50)$$

for $\alpha \ll 1$ and

$$\epsilon_g \approx \frac{0.62}{\alpha}, \quad (51)$$

for $\alpha \gg 1$.

(I hope soon to put some figures here illustrating the $\alpha$-dependence of $\epsilon_g$.)

**discussion**

Now back to the question of how many experiments it takes to determine the genome. The answer is that it isn’t quite as bad as you might fear; that
would correspond to replacing $\alpha$ by $\alpha f$ everywhere in the undiluted solution. We saw that this was the case only for $\alpha \ll 1$. Already for $\alpha = \mathcal{O}(1)$ (i.e. $p = \mathcal{O}(K)$) improvement is significant in the sense that corrections to the small-$\alpha$ limiting form of $q(\alpha)$ (48) are of the same order as (48) itself, and the large-$\alpha$ approach to perfect knowledge is the same as if we knew a priori just which connections were nonzero. At the characteristic crossover value $\alpha_c = \mathcal{O}[\log(1/f^2)]$, improvement of $\mathcal{O}(1)$ occurs.

This value of $\alpha_c$ can be understood crudely in information-theoretic terms. Finding $N$ parameters, given the prior knowledge that all but a fraction $f$ of them are zero, is not as hard as finding them in the absence of this knowledge – the a priori entropy is not as large. Explicitly, the a priori entropy associated with the $n_j$ is

$$S_0 = N \left[ f \log \frac{1}{f} + (1 - f) \log \frac{1}{1 - f} \right] \approx K \log \frac{1}{f},$$  

and this many examples is sufficient to pin them down.

References

