Computing expression patterns from regulatory sequences

04/17/2012
Transcriptional regulation in eukaryotes

- Gene expression patterns in eukaryotes are specified by the interactions between a set of TFs and a cluster of their respective binding sites called a cis-regulatory module (CRM), spanning 500 ~ 2500 bps in the regulatory regions.

- According to its over effect, a CRMs is also called an enhancer, a silencers or an insulator.


INR, Initiator sequences
DPE, downstream promoter elements
Transcriptional regulation in eukaryotes

- The expression rate of a gene in a cell is determined by the strength the interactions between these TFs and their respective binding sites in the related CRMs, which is in turn determined by the following factors:

1. The concentrations of the TFs in the cell;
2. Number of accessible binding sites for each TF in the CRMs
3. Affinity of each TF to each of its binding site in the CRMs;
4. Cooperative interactions among TFs in the CRMs.
Modeling of transcriptional rate in eukaryotes

- One way to compute the expression rate of a gene product is to find its input function.

- As we have shown earlier, if we neglect the cooperative interactions among TFs in the CRM, and assume an additive relationship among the TFs, then, the expression rate of the gene can be coarsely model by a Hill equation (or a logic function),

\[
f(X^*_1, X^*_2, \ldots, X^*_n) = \frac{\sum \beta_i (X^*_i / k_i)^{n_i}}{1 + \sum \beta_i (X^*_i / k_i)^{m_i}}.
\]

- In principle, however, if we know the concentration TFs in a cell and the binding site profile of each TF, we should be able to predict the expression rate of a given gene in the cell based on basic physical properties (thermodynamics) of the transcription initiation process.
Prediction of transcriptional rate in eukaryotes

- This problem can be formulated as fellows,

  Given 1) the concentration profile of $N$ TFs in a cell, 2) the profiles of their binding motifs in the genome, and 3) the regulatory region of a gene, predict the expression rate the gene.

- The concentration of TFs in the cell and available binding sites in the CRMs of the gene determine a **binding configuration** of the TFs to the CRMs.

*Nature Reviews Genetics*; Jul 2009, Vol. 10 Issue 7, p443-456,
Prediction of transcriptional rate in eukaryotes


- Background: During the early embryogenesis of Drosophila, eight maternal TFs (morphogens) form gradients spanning the entire anterior-posterior axis of the syncytium; they are translated into non-periodic domains of zygotic gap genes, and subsequently into periodic patterns of seven “pair rule” genes, and finally 14 segmental stripes that prefigure the fourteen segments of the larva.

EMBO reports 2, 12, 1083–1088 (2001)
Predicting expression patterns in Drosophila embryos during segmentation

- The eight morphogens are:
  - **Activators**: Biocoid (BCD), Torso-response element (TorRE), Caudal (CAD).
  - **Repressors**: Hunchback (HB), Giant (GT), Kruppel (KR), Knirps (KNI), Tailless (TLL)

- The spatial-temporal distribution of relative concentrations of the eight TFs are known;
- A limited number of binding sites of these 8 TFs are also known;
- The spatial-temporal distribution of 44 gap and pair-rule genes are also known.
Overview of the thermodynamic model and approach for predicting gene expression patterns

Input
- Module sequence
- Input factor distribution
- Binding preferences

Thermodynamic modelling
- Binding energy along module sequence, for a given AP position
- Configurations of factor binding to DNA and resulting expression

Output
- Factor occupancy over all AP positions
- Predicted module expression
- Measured module expression

Fit parameters to maximize agreement between measured and predicted patterns
The two components of the algorithm

- Step 1: the sequence component:

  For each possible binding configurations \( c \), compute the probability \( P(c) \) at a position on the anterior-posterior axis.

  A configuration is defined to avoid overlapping between two factors in one configuration. Intuitively, \( P(c) \) is computed based on:

  1. the concentration of TFs at the position;

  2. the strength of the binding of each TF to its binding site.

  3. allowing self-cooperative interaction between the two TF molecules bound to neighboring sites.
The two components of the algorithm

- **Step 2: the translation component:**
  Compute the probability of a gene of having expression level $E$ under the binding configuration $c$, $P(E/c)$;
  Then, the overall expression level is the summation over all configurations,
  \[
  P(E) = \sum_{c \in C} P(c)P(E/c).
  \]
  This probability is actually the probability that RNAP binds the promoter, but it is proportional the transcription rate and the gene product concentration.
Sequence component: distribution of configurations

- A configuration is legal if there is no overlap between two TF molecules on the DNA.

- According to Boltzmann distribution, the probability of each configuration is given by,

\[ P(c) = \frac{W(c)}{\sum_{c' \in C} W(c')} \]

where \( W(c) \) represents the statistical weight associated with configuration \( c \).

- Intuitively, \( W(c) \) should be the product of the contribution of each TFs in \( c \).

- The contribution of a TF is determined by its concentration in the cell and its affinity to the binding site.
Sequence component: distribution of configurations

- Since the concentration of TF \( i \) is usually given as a relative concentration \( R_i \), its absolute concentration can be expressed as,
  \[
  \tau_i = \alpha_i R_i.
  \]

- If a sequence \( S_1, \ldots, S_{L(i)} \) is a binding site of TF \( i \) for the gene, the affinity of \( i \) to the site is,
  \[
  P_i(S_1, \ldots, S_{L(i)}) = \prod_{j=1}^{L(i)} P_i^j(S_j),
  \]
  where \( L(i) \) is the length of the binding site and \( P_i^j(S_j) \) is the score according to the PSWM of the motif.

- If the configuration \( c \) contains binding sites of \( k \) TFs, \( f(i), \ldots f(k) \), then the statistical weight is,
  \[
  W(c) = \prod_{i=1}^{k} \tau_{f(i)} \frac{P_i(S_{p(i)}, \ldots, S_{p(i)+L(i)})}{P_B(S_{p(i)}, \ldots, S_{p(i)\mid L(i)})},
  \]
  where \( P_B \) is the background model distribution, e.g., zero order Markov model.
Sequence component: distribution of configurations

Assume a CRM of a gene contains one and four binding sites of CAD and KNL, respectively, then the weight of the binding configuration is computed as follows if binding cooperativity is not considered:

Configuration Weight $W(c)$

\[
W(c) = \frac{\tau_{kni} P_{kni}(S_{20}, S_{30})}{P_B(S_{20}, S_{30})} \cdot \frac{\tau_{kni} P_{kni}(S_{50}, S_{60})}{P_B(S_{50}, S_{60})} \cdot \frac{\tau_{cad} P_{cad}(S_{200}, S_{208})}{P_B(S_{200}, S_{208})} \cdot \frac{\tau_{kni} P_{kni}(S_{320}, S_{330})}{P_B(S_{320}, S_{330})}
\]
Inccoporating binding cooperativity

- If we consider the self-cooperative interaction between two adjacent binding sites $i$ and $j$ of the same TF, separated by a distance $d$, as $\gamma(i,j,d)$, then the statistical weight becomes,

$$W(c) = \left( \prod_{i=1}^{k} \tau_{f(i)} \frac{P_i(S_{p(i)}, \ldots, S_{p(i)+L(i)})}{P_B(S_{p(i)}, \ldots, S_{p(i)+L(i)})} \right) \left( \prod_{i=1}^{k-1} \gamma(i, i+1, p(i+1) - p(i)) \right)$$

- We assume the strength of cooperatively binding interaction decreases with the distance between the adjacent binding sites, and model $\gamma(i,j,d)$ as a Gaussian function with mean = 0, and standard deviation = 50.
When the cooperativity of adjacent binding site of the same TF is considered, then the weight of the binding configuration is computed as follows:

\[
W(c) = \frac{\tau_{kni}}{P_B(S_{20}, S_{30})} \frac{\tau_{kni}}{P_B(S_{50}, S_{60})} \frac{\tau_{cad}}{P_B(S_{200}, S_{208})} \frac{\tau_{kni}}{P_B(S_{240}, S_{250})} \frac{\tau_{kni}}{P_B(S_{320}, S_{330})} \gamma(kni, kni, 30) \frac{\tau_{kni}}{P_B(S_{320}, S_{330})} \gamma(kni, kni, 80)
\]
Expression component: translating configurations to expression

- Assume $c$ contains binding sites of $k$ TFs: $f(1), \ldots f(k)$, we use logistic function to model $P(E/c)$:

$$P(E \mid c) = \logit \left( w_0 + \sum_{i=1}^{k} w_{f(i)} \right) = \frac{1}{1 + \exp \left( - \left( w_0 + \sum_{i=1}^{k} w_{f(i)} \right) \right)}$$

where $w_0$ is the basal expression level, and $w_{f(i)}$ is the expression contribution of TF $f(i)$.

- To correct the effect of length of the regulatory sequences, the summed contribution of TFs is normalized to the length of the sequence, $L$

$$P(E / c) = \frac{1}{1 + \exp \left( - \left( w_0 + \frac{1}{L} \sum_{i=1}^{k} w_{f(i)} \right) \right)}$$
Put it all together: a unified model over sequence and expression

- The model can be applied to all the locations along the anterior-posterior axis to compute the expression pattern of a gene along the axis.
- The only differences is the concentration of each TF at a specific location.
Optimizing the model parameters

- This model contains a few parameters to be estimated using a training dataset before it becomes a real expression pattern predictor:
  1. The concentration scaling factor $\alpha_i$ for TF $f(i)$, which is used to compute the absolute concentration of $f(i)$,
     \[ \tau_i = \alpha_i R_i ; \]
  2. Self-cooperativity binding strength,
     \[ \gamma(i, j, d) ; \]
  3. Expression contribution weight $w_i$ of each TF $f(i)$.

- The parameters can be estimated by minimizing the differences between the predicted expression levels of all genes and their measured level along all positions of the spatial axis in the training dataset.
Optimizing the model parameters

To describe the training procedure, let’s define the following variables:

- $S$: a set of regulatory sequences associated with a set of genes,
- $A$: a spatial axis,
- $P(E/S,A)$: predicted expression along $A$ of the set of genes, and
- $O(S,A)$: measured/observed expression of the genes.

Then, the total squared difference between the predicted and measured expression level across all input sequences and positions of the spatial axis is given by,

$$F = \sum_{s \in S} \sum_{a \in A} (P(E/s,a) - O(s,a))^2.$$ 

This is the **objective function** that we want to minimize, and our goal is to find the sets of parameters $\alpha = \{\alpha_1, \ldots, \alpha_k\}$, $\gamma = \{\gamma_1, \ldots, \gamma_k\}$, and $w = \{w_1, \ldots, w_k\}$, which minimize $F$. 
Optimizing the model parameters

- Two different standard minimization procedures can be alternatively applied to minimize the objective function to avoid being stuck in a local minima.
  1. Conjugate gradient ascent method;
  2. Simplex method.

- In order to apply these methods, we first need to compute the expression probability of each sequence $s$ at position $a$,

$$P(E / s, a) = \sum_{c \in C} P(c / s, a)P(E / c, s, a).$$

- Given a configuration $c$, its expression contribution is a logistic function and is easy to compute,

$$P(E / c, s, a) = P(E / c, a) = \frac{1}{1 + \exp\left(-\left(w_0 + \frac{1}{L} \sum_{i=1}^{k} w_{f(i)}\right)\right)}.$$
Optimizing the model parameters

- The probability of each configuration is given by,

\[ P(c / s, a) = \frac{W(c / s, a)}{\sum_{c' \in C} W(c' / s, a)}. \]

- Since the number of configurations, \(|C|\), increases exponentially with the number of TFs in C, the brute-force solution to compute the denominator will not work.

- Let’s consider a simplified case:

Given a sequence of length \(L\), if we have \(M\) different TFs, and there is a binding site every \(N\) bases (assume that \(L\) is some integer units of \(N\)), then there are \(L/N\) binding sites, and the number of possible configuration is \(M^{L/N}\).

If \(M=10\), \(L=1000\), \(N=50\), then \(M^{L/N} = 10^{1000/50} = 10^{20}\)