Tandem Promoter Strength Prediction Model.(Abbreviation: TP Model)

Content:

1. Overview

This model aims at predicting the final output of a tandem promoter system, which can be constituted of any number of and any type of sub-promoter(including sub-tandem promoter) in any order and any species. The Key idea of the model is that the strength of a promoter system is proportional to the probability of at least one RNA Polymerase (mentioned as RNAP latter) binding on the promoter.

2. Symbol table, Assumption and reasons.

Symbol	
[ ]	The symbol of concentration, i.e. [Protein] means the concentration of the protein
p_{tot} / y	The probability of at least one RNAP(with all of its subunit) binding on the tandem promoter, also represents the normalized strength of the promoter.
n / x	The number of sub-promoters in the tandem promoter system.
u	Number of copies of a tandem promoter in a cell
ξ	Strength constant, equals to the strongest expression level possible (units in fluorenes normalized by a internal reference).
V	The volume of a cell
p_i	The probability of a RNAP(with all of its subunit) form a RNAP-with complex with the i^{th} sub-promoter in the tandem promoter system.
q_i	$q_i=1-p_i$, the probability of a RNAP not binding to the i^{th} sub-promoter
j	Cooperative factor
α	Transcription rate constant
λ	mRNA degradation constant
v	Translation rate constant
k	Protein degradation constant
Abbreviation	
RNAP	RNA Polymerase
ODE	Ordinary Differential Equation
RP / RP_c	RNAP-Promoter complex, inactive complex
RP_i	Intermediate complex
RP_o	Open complex

Table 1. Symbol table of TP Model

1.It's assumed that the promoter strength is measured in the same species, with identical environment and growing stage. This ensure the assumption that the concentration of all subunits of RNAP, all subunits of ribosome, all RNA degradation enzymes, all kind of proteases and all transportation protein are thermodynamically

identical. Otherwise, the model may fail to work properly.
2.In all measurement, the contexts of the promoter are the same. i.e. same RBS, terminator, protein sequence, up stream element, down stream element and DNA supercoiling.
3.All transcriptional factors are not considered in this version of the model, but can be included in the model with some modification to the equations.
4.The promoter region is accessible for RNAP(and all kinds of its subunits), which means it's not in heterochromatin region or any other condition that hamper a normal RNAP-DNA interaction.
5.The probability of RNAP binding on the region between two sub-promoter within the tandem promoter system is neglected. As it contributes too little to final p_{tot}.
6.The RNAP-DNA binding is assumed to stay on equilibrium in the model. This is reasonable because the open complex formation is a slow rate limiting step of transcription. So in the time scale of open complex formation, RNAP-DNA binding can always reach its equilibrium in neglectable time[1][2]. It's also observed that the inactive RNAP-DNA complex can be detected on the DNA[3].

3. Modeling result

We found that the strength of a tandem promoter system can be interpreted by a simple equation:

$$Strength = \frac{u\xi}{V}[1 - \prod_{i}^{n}(1 - p_i n^j)] \quad (1)$$

Where q_i is the probability of a RNAP(with all of its subunit) not forming a RNAP-with complex with the i^{th} sub-promoter, n the number of sub-promoters, j the coordinative factor, and ξ the strength constant.

If we define the highest possible expression level of a promoter in certain species is 1. Then the equation 1 become normalized.

$$Strength' = \frac{Strength}{Strength_{\max}} = p_{tot} = 1 - \prod_{i}^{n}(1 - p_i n^j) \quad (2)$$

This model explains 99% of the tandem promoter strength variation caused by
1.number of sub-promoter,
2.kind of sub-promoter,
3.order of sub-promoter .
(With a R-square=0.992 and confidence bond of 95% when fitted with our data)

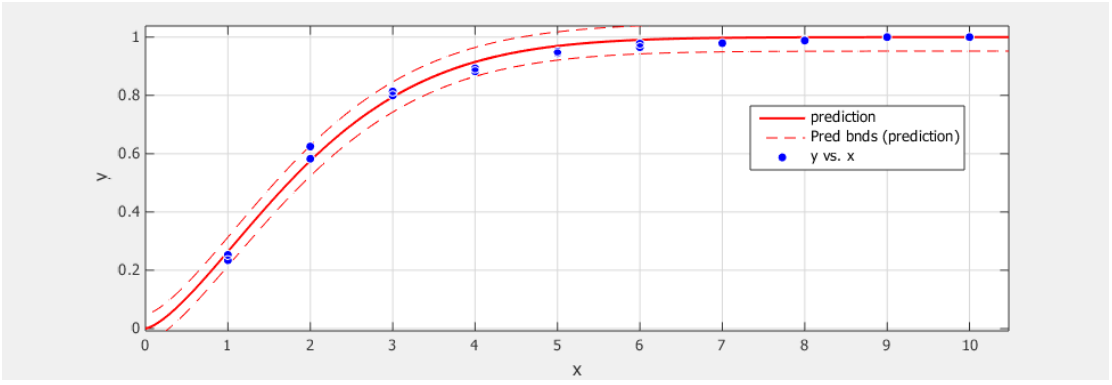

Figure 1. Model fitting result
Y-axis represent the normalized promoter strength, X-axis the number of sub-promoter

The blue dot is data extracted from ref.[4] fig.2, the red line is the prediction made by our model, the red dotted line is the 95% prediction bound

4.Model derivation

The promoter strength may be influenced by various factors. We need to simplify the system into some reasonable toy model by wiping out all relatively trivial factor.

4.1 Expression level Measurement

We use the fluorescence strength to indicate the strength of the promoter(Normalized by a inner reference fluorescence protein(FP) - mCherry. Please check details at the experiment part). Because when the exciting light is fixed, the fluorescence is proportional to the concentration of FP. And FP can be lighted up in a short time after they are synthesis.

4.2 Translation and transcription

According to the Central Dogma.

$$\text{DNA} \rightarrow \text{RNA} \rightarrow \text{Protein}$$

So we can write down the following ODE, which is similar to the equations in [5].

$$\frac{d[mRNA]}{dt} = \alpha[RP] - \lambda[mRNA] \quad (3)$$

$$\frac{d[protein]}{dt} = v[mRNA] - k[protein] \quad (4)$$

Where α means the mRNA producing constant, λ the mRNA degradation constant, v the protein synthesizing, k the protein degradation constant, and [RP] is the concentration of RNAP-promoter complex.

In equation 4, the protein increasing speed is determined by [mRNA] and v. With same RBS, v relates to the efficiency and concentration of ribosome and concentration of amino acids in the cell, which can be considered identical under the experiment condition of comparing different promoter. The protein degradation speed is determined by [protein] and k. k relates to protease system in the cell, which can also be considered as identical in measurements between different promoter.

In equation 3, the mRNA increasing speed is determined by [RP] and α, and its degradation depends on [mRNA] and λ. Both α and λ can be treated as constant in the experimental condition of comparing different promoter. As α depends on the transcription initiation efficiency, which is assumed to be identical for any RNAP-DNA complex at this stage, for simplicity. This is reasonable because if α varies, the difference of α can be incorporated in [RP] (and finally in p_i, see latter derivation). Though this part of the equation varies from the equations in [5], it is justified by the phenomenon that when [RNAP] and [DNA] is hold in a constant, the UTP incorporation is a zero order reaction [2]. And λ depends on the concentration of RNase which doesn't varies in different promoter measurement.

Therefore, because we are interested in the steady state of the protein expression. We can set,

$$\frac{d[mRNA]}{dt} = \frac{d[protein]}{dt} = 0$$

$$\therefore [protein]_{eq} = \frac{\nu\alpha}{\lambda k}[RP]$$

We can consider [protein]eq as the indicator of the promoter strength, and let να/ λk=ξ

$$\therefore Strength = \frac{\nu\alpha}{\lambda k}[RP] = \xi[RP] \quad (5)$$

So the strength of the promoter is directly related to the concentration of the RNAP-DNA complex of this promoter.

4.3 RNAP binding and transcription initiation

The open complex formation reaction is as follow.

$$DNA + RNAP \underset{}{\overset{K_1}{\rightleftarrows}} RP_c \underset{k_1}{\overset{k_2}{\rightleftarrows}} RP_i \underset{k_{-3}}{\overset{k_3}{\rightleftarrows}} RP_o$$

Where RP_c is the inactive complex, RP_i is the intermediate complex and RP_o the open complex.

The reaction can be combined with Central Dogma to be:

$$DNA + RNAP \xrightarrow{K_1} RP \xrightarrow{K_{slow}} protein$$

Because K_1 happens in a much smaller time scale. The probability of finding the polymerase on the promoter will be given by its equilibrium constant K_1.[1]

To evaluate the probability of polymerase binding (p_i) we must sum the Boltzmann weights over all possible states of P polymerase molecules on DNA.

$$Z(P) = \underbrace{\frac{N!}{P!(N-P)!}}_{number\ of\ arrangements} \times \underbrace{e^{-\frac{P\varepsilon^{NS}}{k_bT}}}_{Boltzmann\ weight}$$

This equation calculate the total Boltzmann weight of no RNAP binding to the target promoter, with N represent the number of non-specific sites on the DNA, P the effective RNAP number, ε^{NS} the non-specific binding energy, kb the Boltzmann constant and T the temperature.

$$Z(\mathrm{P}-1)\,Z_i = Z(\mathrm{P}-1)e^{-\frac{\varepsilon^{\mathrm{Si}}}{k_bT}}$$

This equation calculate the total Boltzmann weight of one RNAP binding to promoter i, with ε^{Si} means the specific binding energy of promoter i.

So the probability of a RNAP binding to promoter i is,

$$p_i = \frac{Z(P-1)Z_i}{Z_{tot}}$$

With Ztot represent the sum of all Boltzmann weight of all different condition.

So the probability of RNAP binding to both promoter i and j is,

$$p_{ij} = \frac{Z(P-2)Z_i Z_j}{Z_{tot}}$$

$$p_i p_j = \frac{Z(P-1)^2 Z_i Z_j}{Z_{tot}^{\ 2}}$$

When $N \gg P \gg 1$, we have $Z_{tot} \approx Z(P)$

$$\frac{p_{ij}}{p_i p_j} = \frac{Z(P-2)Z_{tot}}{Z(P-1)^2} = \frac{\frac{N!}{(P-2)!(N-P+2)!} \times \frac{N!}{P!(N-P)!}}{(\frac{N!}{(P-1)!(N-P+1)!})^2} = \frac{(N-P+1)(P-1)}{(N-P+2)P} = \frac{NP}{NP} = 1$$

So the probability of RNAP binding to two promoter at the same time equals to the product of the probabilities of RNAP binding to the two promoter respectively. i.e.

$p_{ij} = p_i p_j$

As only one RNAP is needed to initiate the transcription in a tandem promoter system (the other RNAP will be blocked by the RNAP closest to the transcription initiation point). So the probability of at least one RNAP binding to the promoter is

$$q_i = 1 - p_i; \quad p_{tot} = 1 - \prod_i^n q_i \quad (6)$$

For a promoter with u copies in a cell (all separated and function independently)

$$[RP] = \frac{up_{tot}}{V} \quad (7)$$

The strength of a promoter is, according to equation 5.

$$Strength = \xi[RP] = \xi \frac{up_{tot}}{V}$$

the maximum strength possible can be reached when p_{tot}=1,

$$Strength_{\max} = \frac{u\xi}{V}$$

$$\therefore Strength' = \frac{Strength}{Strength_{\max}} = p_{tot} = 1 - \prod_i^n q_i \quad (8)$$

However, we found this model can not fully explain our data. The fitting result, though has a satisfactory R-square(0.948), fail to explain the great difference between our model prediction and the data when there's only one promoter in the "tandem promoter system". This means that the p_i we found by curve fitting is not the real p_i.

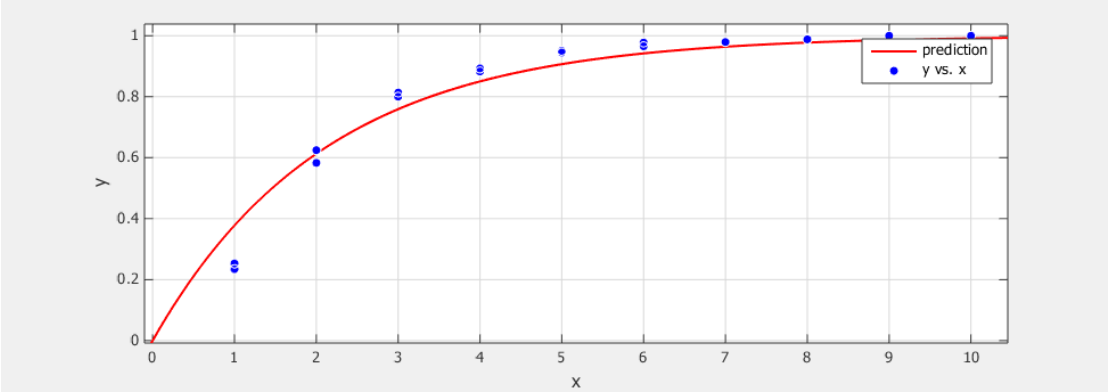

Figure 2. Model fitting result of the simpler model

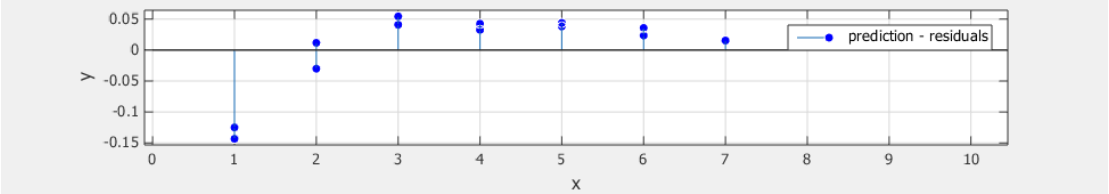

Figure 3. Curve fitting residual plot of the simpler model

Data analysis shows that the data increase in y much quicker than our prediction, which indicate there will be some kind of cooperation among sub-promoters. This results in $p_{ij}>p_ip_j$. The cooperation can be explained by the fact that when one RP_o formed, it will "melt" the DNA duplex into two single strain. This DNA untwisting, unwinding and melting make the RNAP-DNA complex in the vicinity easier to transform from RP_c to RP_o. Therefore variation in α can no longer be ignored.

Now consider the situation when variance in α is considered. Set α=α*var(α), where α* is the standard "α", and var(α) is the degree of α varies from α*. Because even considering the possible variance in α, the transcription initiation is still much slower than RNAP-DNA binding[6]. The time-scale separation is still valid (the RNAP-DNA can still be considered in equilibrium).

According to equation 5 and 7.

$$\therefore Strength = \frac{v\alpha}{\lambda k}\frac{up_{tot}}{V} = \frac{vu\alpha^*}{\lambda kV}\mathrm{var}(\alpha)p_{tot}$$

We can incorporate var(α) into p. Actually, as we get our data of p_i from fluorescence experiment. The var(α) of different protein has already incorporated into p_i.

But the cooperative var(α) hasn't been incorporated to any p_i. So we should add a adjust term(the cooperation factor) into equation 8. Therefore equation 2 comes out, with n^j as the cooperative factor.

$$Strength' = \frac{Strength}{Strength_{\max}} = p_{tot} = 1 - \prod_i^n (1 - p_i n^j) \quad (2)$$

As we've showed in figure 1. This model successfully captures the essence of tandem promoter system. With the residual plot as follow.

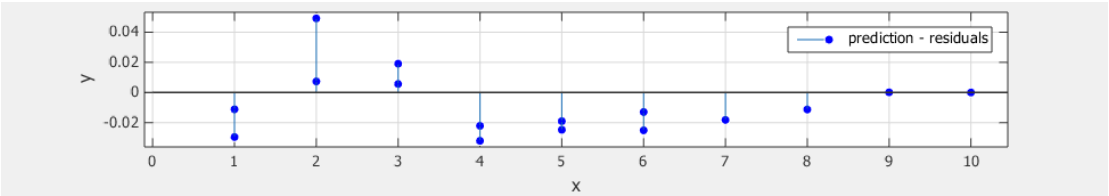

Figure 4. Curve fitting residual plot of the final model

This model isn't flawless, as the cooperative factor just fitted into the data well but has no solid biological ground(it's even a boundless function when x approach infinite). The more prudent way will be choosing a sigmoid function rather than n^j. But that will make the model to complex, and hard to employ when people just have scarce data about their promoter (easy over-fitting). So we decide to keep it in this simpler and efficient form.

5.User Guideline

To employ the model, the user need to assign the p_i for each kind of promoter that will be used to construct the tandem promoter.

The simplest way to achieve it is as follow.
1)Using fluorescence protein to indicate the expression level of each promoter or promoter association, optional (normalize it by a internal reference just as we used a RFP in our experiment).

2)To measure the strongest expression level possible in the species. Using a known strongest promoter to construct a tandem promoter that made of 5 repeats of the promoter, to see the strongest expression level.

3)Normalizing other promoter's expression level by the strongest expression level, which result in the p_i of each promoter. As follow.

$$p_{tot} = \frac{V \cdot Strength}{\xi u}$$

4)using equation 2 to predict the p_{tot} of the designed tandem promoter, with an empirical cooperative factor j=0.4.

$$Strength' = \frac{Strength}{Strength_{\max}} = p_{tot} = 1 - \prod_{i}^{n} (1 - p_i n^j) \quad (2)$$

In this way, the error of the prediction should be less than 4% of the maximum expression rate, as our data showed before.

If the data allow, the user can carry out fit with a variable j, which may varies in different species and cell condition.

Reference:
1.Bintu, Lacramioara, et al. "Transcriptional regulation by the numbers: models." *Current opinion in genetics & development* 15.2 (2005): 116-124.
2.Buc, Henri, and William R. McClure. "Kinetics of open complex formation between Escherichia coli RNA polymerase and the lac UV5 promoter. Evidence for a

sequential mechanism involving three steps." *Biochemistry* 24.11 (1985): 2712-2723.
3.DeHaseth, Pieter L., and John D. Helmann. "Open complex formation by Escherichia coli RNA polymerase: the mechanism of polymerase‐induced strand separation of double helical DNA." *Molecular microbiology* 16.5 (1995): 817-824.
4.Li, Mingji, et al. "A strategy of gene overexpression based on tandem repetitive promoters in Escherichia coli." *Microbial Cell Factory* 11 (2012): 19.
5.Buchler, Nicolas E., Ulrich Gerland, and Terence Hwa. "Nonlinear protein degradation and the function of genetic circuits." *Proceedings of the National Academy of Sciences of the United States of America* 102.27 (2005): 9559-9564.
6.Alon, Uri. *Introduction to Systems Biology: And the Design Principles of Biological Networks*. Vol. 10. CRC press, 2007. Page 6.