Machine Learning in Bioinformatics – Gene Regulation Network

التعلم الآلي في المعلوماتية الحيوية ـ شبكة تنظيم الجينات

Sameerah Faris Khlebus
University of Information Technology and Communications, College of Business Informatics

Abstract

A biological cell is a complex and complicated environment, where thousands the entities interact surprisingly between each other. This the integrated device the continuously receives external and internal signals to carry out the most vital processes to sustain life. Although thousands the interactions are stimulated in very small areas, biologists assert that, there are no the collisions or the incidental events. On other the hand, rapid discoveries in biology and the rapid evolution of data collection make it difficult to build a concrete perspective that scientifically explains all observations. Cooperation has therefore become necessary among physicists, mathematicians, biologists and the computer engineers. The aim of this virtual company is to pursue what is known as biological network modelling.

Keywords: Bioinformatics, Machine learning, Biochemistry, System biology, stochastic system.

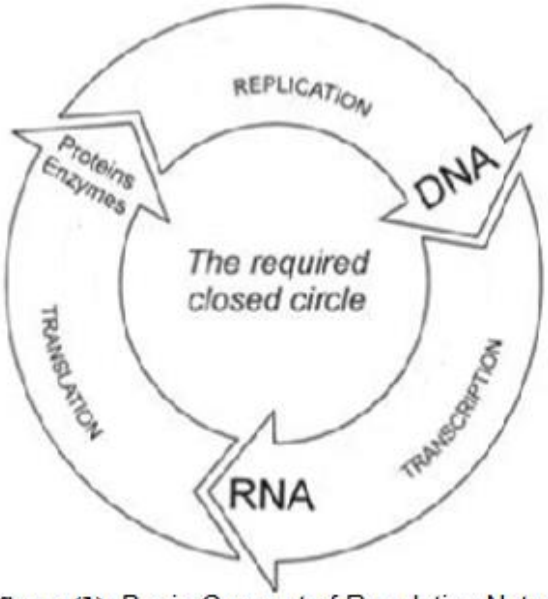
ملخص البحث

الخلية بيولوجية هي بيئة معقدة، والآلاف من الكيانات تتفاعل بشكل مفاجئ بين بعضها البعض. هذا الجهاز المتكامل يتلقى باستمرار الإشارات الداخلية والخارجية لإجراء معظم العمليات الحيوية للحفاظ على استمرار الحياة. على الرغم من تحفيز الآلاف من التفاعلات في مساحات صغيرة جدا، وعلماء الأحياء يؤكدون عدم وجود الصدف أو أحداث عرضية. من ناحية أخرى، والاكتشافات السريعة في مجال البيولوجيا والتطور السريع في تجمع البيانات تجعل الأمر أكثر صعوبة لبناء منظور ملموس على أن يفسر علميا جميع الملاحظات. وبالتالي، أصبح التعاون الضروري بين الأحياء، الرياضيات، الفيزياء ومهندسين الكمبيوتر. والهدف من هذه الشركة الافتراضية هو عمل وتحقيق ما يعرف بشبكة النمذجة البيولوجية.

الكلمات الدالة : المعلوماتية الحيوية، تعلم الآلة، الكيمياء الحيوية، وعلم الأحياء ، نظام مؤشر stochastic.

1. Introduction

Modeling of Gene Regulation Network a cell is a complicated (having a huge number of elements) and complex (behaving surprisingly) environment. Most vital processes, those are necessary for any organism's life, happen inside this integrated device. A cell continuously receives internal and external signals for producing proteins, also known as gene products. These components perform the main functions for a cell's life. Also, a cell controls its environment in a way that determines the rate at which each protein is needed. Within the cell the operation of information processing is known as transcriptional network ,figure (1). A transcriptional network, whose end is to synthesize proteins in a process called translation, consists of many layers. The first layer in which mRNA is produced is known as transcription process. Firstly, our focus, in the following sections, will be on the general principles of constructing the gene regulation network. For clarity, we have suggested a framework of how to build this kind of model containing the main steps. For sure, the details, within each process of model construction, vary depending on, for instance experiment's type, the tools used to analyze the data and the software used to simulate the suggested model. Secondly, in the later sections, we expose to one computational method that was developed to reconstruct dynamic regulatory maps known as Dynamic Regulatory Events Miner (DREM) [2].



figure(1): Basic Concept of Regulation Natwork

1.1 Complexity of system biology

Any system is a set of components that interact with each other to achieve the final aim. More sophisticated systems consist of sub-systems, which jointly perform more complicated tasks. Then the whole network may be connected to some other network for more tasks accomplishment, and so on. This is exactly the case in biological system. For instance, consider two TFs that interact, independently or cooperatively, to catalyze the RNAp to bind the promoter for protein production. Further process would, for example, comprise interactions between amino acids of the produced protein that may bind to the binding site to act as repressor to its own production. Clearly the system can be defined at different hierarchical levels and this will determine the extent of detailed at which the components of the system should be described usefully. For instance, when considering a protein as a system one could look at the amino acids level while someone studying glycolysis would typically look at the protein level. Thus, the biological cell has, in fact, remarkable capabilities:

It catalyzes, in a coordinated fashion, thousands of reactions in very small spaces.

- These reactions include copying the whole cellular machinery to make another cell, cell cycle.
- The cell has a variable make-up enabling it to adapt to varying conditions as encoded in the DNA.
- The base sequence in the DNA is a subject to mutation which change the blueprint, and thereby, the cellular composition [1, p. 14].

1.2 Framework of Modeling

The biological cell is a very complex environment, and thereby, capturing all reactions in one model would be computationally expensive and waste of time. Additionally, the discovery in this field going very fast and any comprehensive model may not be correct due to the new data, and hence the parameters and parameters values of the model are not correct. Good approach is to divide the whole system into subsystems, then after test each one to investigate the correctness and lastly integrate all these sub-systems to have the big one. The following pathway, shown in figure (2), represents the main steps of constructing a model of system biology.

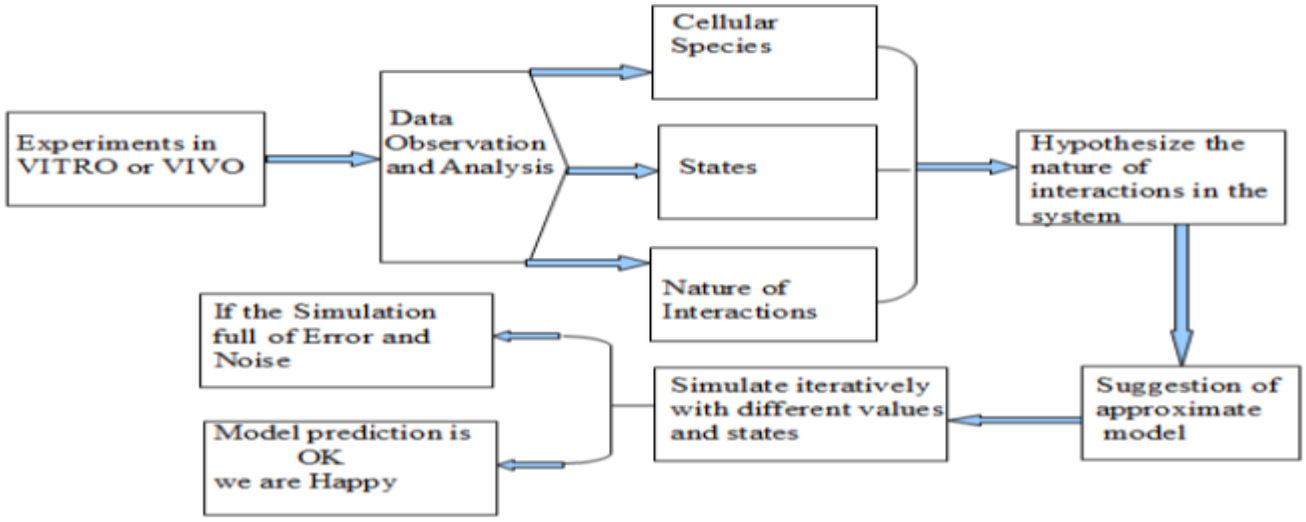


Figure (2): The framework of constructing a biological system model.

Information one can get in the first three stages are not necessarily obtained by doing the assays in VITRO or VIVO, but one can get this information by reading literature, searching databases and investigating the studies of particular system. Several data sets are available such as ChIP-chip, motif, static expression experiments and time-series expression experiments data. These data sets provided by biologists are the corner stone in constructing transcriptional network's models. Various computational methods are used as data mining engine to search and integrate information regarding gene regulatory network. In other words, modeling can be jointly done by biologists and bioinformatics each have specific role. This information can be analyzed and processed in such a way to put them in logical and temporal sequence. The consequent of these stages should answer questions like; how many cellular species (reactants and products) are involved? What are the initial values of each species? What are the elementary reactions produced by these species? What is the stochastic rate constant of each reaction? And so on. Then, next step is coming.

1.2.1 Hypothesize the nature of interactions in the system

The best way to generalize the nature of interactions in the biochemical is by utilizing mathematically a kinetic model. Many approaches are used to describe the nature of interactions of cellular elements of the biological system such as ordinary differential equations (ODEs) and Gillespie algorithm.

Ordinary Differential Equations (ODEs): This model manipulates the nature of interactions of the system mathematically. A useful form that describe the effect of transcriptional factor on the transcriptional rate of many real gene input function is an increasing S-shape function called Hill Function [2, p.13].

$$f(x^*) = \frac{\beta X^{*n}}{K^n + X^{*n}} \quad \text{... Hill Function for activator}$$

Where K is activation coefficient and has unit of concentration, B is expression level of promoter and n represent Hill coefficient which determines the steepness of the function. On the other hand, for repressor, the Hill input function is a decreasing S-shape takes the form:

$$f(x) = \frac{\beta}{1+\left(\frac{X^*}{K}\right)^n} \quad \text{... Hill Function for repressor}$$

Figure (3) illustrates the effect of Hill coefficient on the shape of the function, the higher the coefficient the steeper the function. Another factor that determines the position of the interaction in time is the ratio of (X/K). That is, can use this factor to shift the whole window to the right or to the left to construct a temporal interaction. By these tow factors the designer can determine the speed of interaction i.e. reach the maximum concentration and the time slot of the interaction, i.e. build cascading interactions.

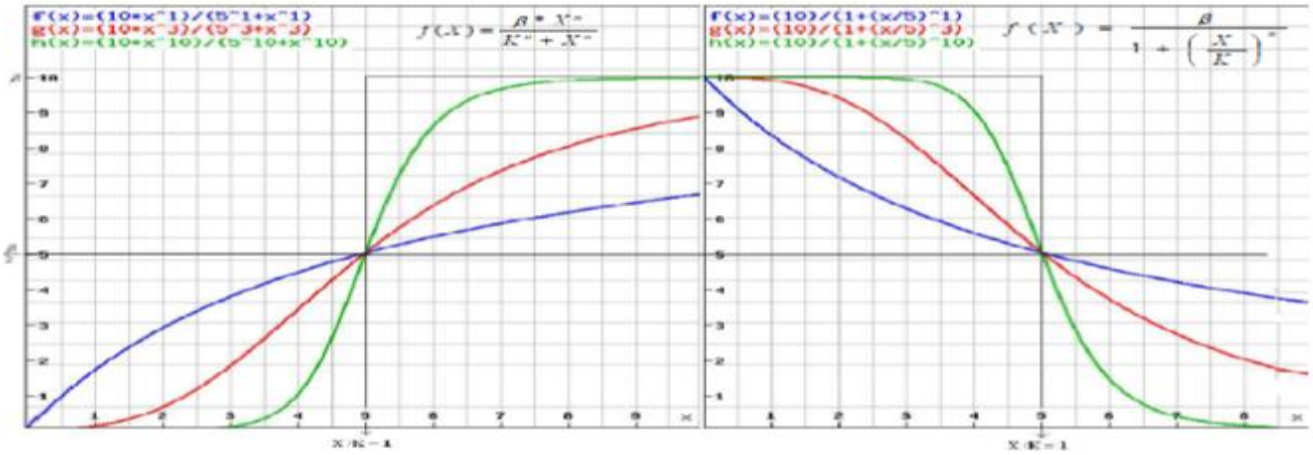


Figure (3): Hill Function with three **different values of** Hill coefficient to prove the steepness effect **[3]**

For instance, consider a gene (DNA stretch) regulated cooperatively by two TFs (A and B) [4, p.79], shown in figure (4).

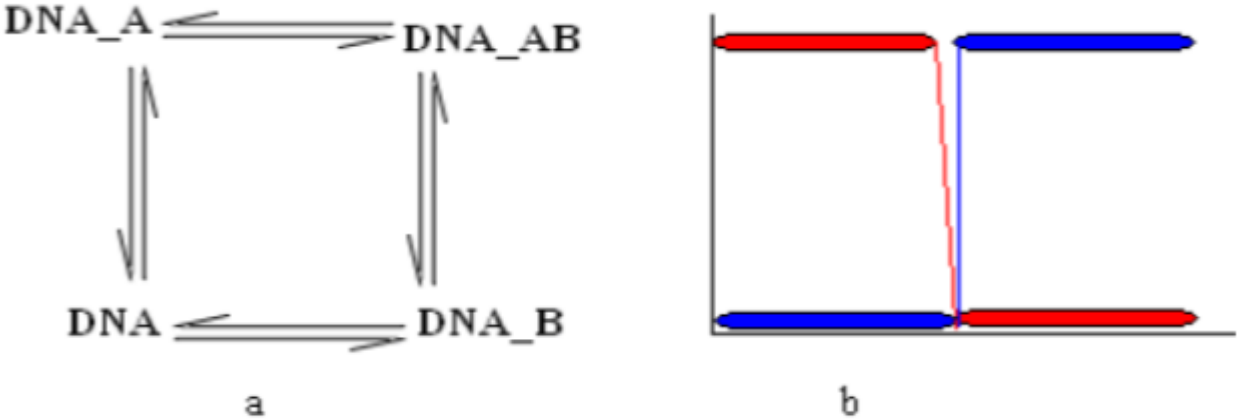


Figure (4): **a) tow TFS regulate cooperatively one. b)simulation this regulation by using step function**

Here are some facts in this transcriptional regulation:

- Both TFs A and B bind to their respective sites (DNA_AB state) for gene regulation to be started. For activation, the logic input function can be described using step- function θ that makes a step when A and B exceed the threshold (KA) and (KB) for (n→ ∞) in the corresponding Hill function.

$$f(A, B) = \beta\ \theta(A > K\,A)\ \theta\ (B > K\,B) \sim A \text{ AND } B.$$

- This regulation represents a binary variable, which is considered as AND gate [2, p16]. Thus the expression level (β) is either zero, when θ = 0, or maximum, when θ = 1. Here DNA = 1 (red) in the first period, while DNA_AB = 0 (blue) and vice versa for the next period where A and B are The kinetic rate is arbitrary, so is the time unit.
- To get DNA_AB complex alternatively two pathways of reverse and elementary reactions forward:

A_ Binds _ DNA,	*DNA+A*	→	*DNA _ A,*	K_{fA}
B_ Binds _DNA A,	*DNA _ A+B*	→	*DNA _ AB,*	K_{fA2AB}
DNA _ AB _ release B,	*DNA _ AB*	→	*DNA _ A+B,*	K_{rAB2A}
DNA _ A _dissociate,	*DNA _ A*	→	*DNA + A,*	K_{rA}
OR				
B _ Binds _ DNA,	*DNA + B*	→	*DNA _ B,*	K_{fB}
A _ Binds _ DNA_ A,	*DNA _ B+A*	→	*DNA _ AB,*	K_{fB2AB}
DNA _ AB _ release A,	*DNA _ AB*	→	*DNA _ B+A,*	K_{rAB2B}
DNA _ B _ dissociate,	*DNA _ B*	→	*DNA + B,*	K_{rB}

Then should calculate is probability of promoter occupancy as a ratio of the frequency of being in state DNA _ AB to the sum of frequencies of all the states on the way to DNA _ AB such as:

$$promoter\ occupancy = \frac{DNA \cdot K_A \cdot K_{A2AB} \cdot A \cdot B}{DNA + DNA \cdot K_A \cdot A + DNA \cdot K_B \cdot B + DNA \cdot K_A \cdot K_{A2AB} \cdot A \cdot B}$$

OR

$$promoter\ occupancy = \frac{DNA \cdot K_B \cdot K_{B2AB} \cdot B \cdot A}{DNA + DNA \cdot K_A \cdot A + DNA \cdot K_B \cdot B + DNA \cdot K_B \cdot K_{B2AB} \cdot B \cdot A}$$

Employing the above facts, taking the account of characteristic of each interaction, can generalize any interaction involved in a particular biological system. This gives the base of proposing a model for further process.

Gillespie algorithm: is the allows only discrete stochastic, simulation of a system with few reactants, because of every reaction is explicitly simulated [5]. When simulated , algorithm realization represents a random walk that represents exactly distribution of the master equation . It is assumed that collisions are frequent, but collisions with the proper orientation and energy are infrequent. So as, all reactions within the Gillespie framework must involve at most two molecules. Reactions involving three molecules are assumed to be extremely rare and are modeled as a sequence of binary reactions. It is also assumed that the reaction environment is well mixed. Gillespie developed two different, but equivalent formulations direct method and first reaction method. The following steps illustrate the operation of the algorithm :

- Initialization: Initialize all numbers of molecules in system, random numbers generators reactions and constants.
- Monte Carlo step: Generate random numbers to determine next reaction to occur as well as the time interval. Probability of a given reaction to be chosen is proportional to the number of substrate molecules.
- Update: Increase time step by the randomly generated time in step 2. Update molecule count based on the reaction that occurred.
- Iterate: Go back to 'Step 1 unless the number of reactants is zero or simulation time has been exceeded [6].

1.2.2 Suggestion of approximate model

Biological systems behave surprisingly, as they are very complex environment, depending on their states. Therefore, the experimentally observed data do not provide a full fathom of a specific system. The reason is that, probably, the assays do not cover all the states that a particular system has in reality. Moreover, in most cases the biological systems are nonlinear, which mean, the designer should be careful about setting the parameters of model components. On the other hand, modeling can help researchers to bridge the gaps in understanding any system. Let us consider the very simple auto-regulatory gene network of prokaryotic shown in figure (5) [7, p. 12] and try to suggest a model assuming that we have done all preceding stages. In the system below, the dimers of produced protein (**P**) by gene (**g**) repress their own transcription by binding to a regulatory region (**q**) upstream of (**g**) and downstream of promoter (**P**). For this biological system we can propose a full and detailed model of all the chemical reactions including RNAp binding,
Ribosome binding to mRNA to produce protein in the cytoplasm, RNase binding to mRNA for degradation and so on.

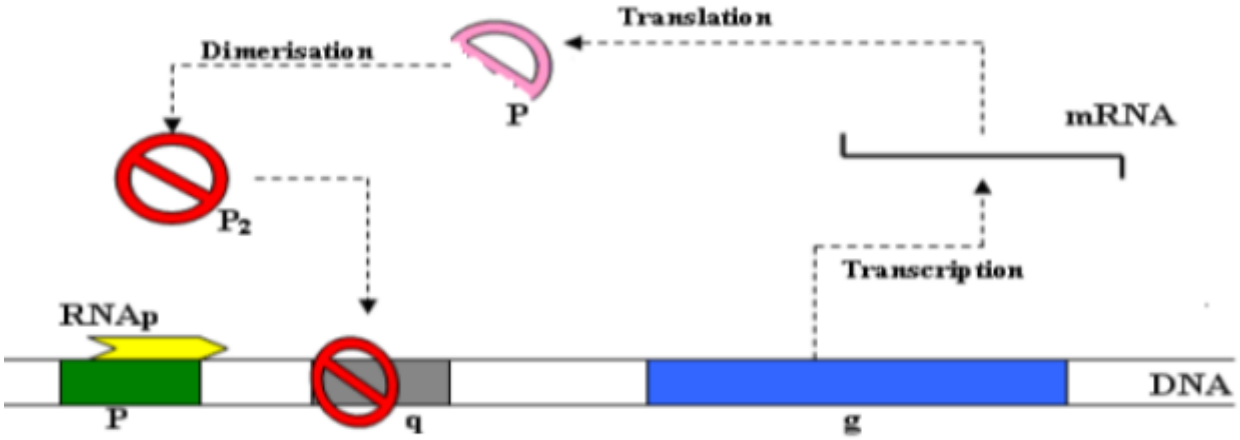


Figure (5): prokaryotic auto-regulatory gene network

For chemical reactions the following notations will be adopted:

$g + P_2$
$\leftrightarrow gP_2$ $repression(+reation\ of\ specis)(\leftrightarrow reversible\ reation)(\cdot combination\ comlex)$
$g + P_2 \leftrightarrow gP_2$ $tramscription(+reation\ result)(one\ dirdction\ reation)$
$g \rightarrow g + mRNA$ $repression(+reation\ result)(one\ dirdction\ reation)$
$mRNA \rightarrow mRNA + P\ translation$
$2P \leftrightarrow P2$ $dimerization$
$mRNA \rightarrow$ $mRNA\ degradation$
$P \rightarrow$ $P\ degradation$

$P_2 + g \leftrightarrow g.P_2$ repression

$g + RNAp \rightarrow g.RNAp$

$g.RNAp \rightarrow g + RNAp + mRNA$ tramscription

$mRNA + Rib \leftrightarrow mRNA.Rib$ translation

$mRNA.Rib \leftrightarrow mRNA + Rib + P$

$P_u \rightarrow P_2$ un folded protein P_u became folded protein P_f

$2P \leftrightarrow P_2$ Dimerization

$mRNA + RNas \leftrightarrow mRNA.RMas$

$mRNA.RNas \leftrightarrow RMas + \emptyset$ mRNA degradation

$P + t \rightarrow P.t$

$P.t \rightarrow t + \emptyset$ Protein degradation by reacting with cell signaling molecule

The schema above shows the whole process as a set of interactions, which start by binding RNAp to promoter to produce an mRNA (transcription). Then in the cytoplasm, Rib will bind to the mRNA to produce a protein (translation). When the folded protein reaches the required threshold, it will perform dimerization and bind to the promoter to act as repressor to its own production. Whereas, mRNA and protein degradations are going on separately.

Also, can representation this model graphically by utilizing Petri net style diagram. This representation is one way to understand the system interactions by explaining themas

a pathway diagram. Some people consider graphical representation is better understanding than a list of reactions mentioned above. By looking at figure (6). Worth noting that, the biological systems, by its very nature, contain “loops” in the reaction network.

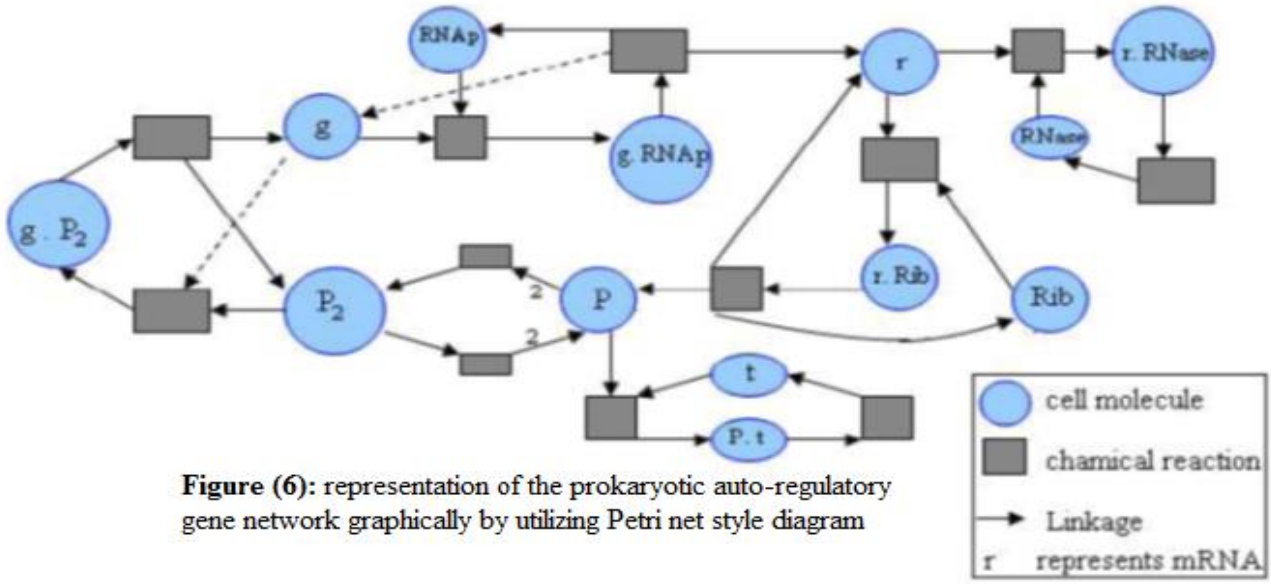


Figure (6): representation of the prokaryotic auto-regulatory gene network graphically by utilizing Petri net style diagram

1.2.3 Simulate iteratively with different values under different conditions

In this last stage, and after the user has chosen the desired software tool to verify his/her findings, it is important realize some facts. For clarity they are going to be listed below:

- Known unknowns: represent the constant variables during the experiment, which should be varied one at the time to check correctness of the model prediction.
- Unknown unknowns : some factors or the interactions that are not aware.
- Validation: this ensures that system has captured true system characteristics rather than experimental noise. In practice, data divided in to 3 sets, testing, training and the validation. Training data comprises data with which model is built [4, p.23].
- For the example above can simulate the interactions to check the model prediction as follows:

1. Simulate the reactions needed for forming the complex that activates the transcription (promoter occupancy).
2. When the full regulatory complex is the formed on promoter, simulate reactions needed for binding RNAp II complex and the initiate transcription with certain probability.
3. Simulate the reactions needed for clearing the promoter region and forming mRNA.
4. Simulate reactions needed for the alternative splicing, RNA editing and RNA regulation.
5. Simulate the reactions needed for mRNA degradation otherwise, mRNA will just keep accumulating.
6. Simulate the reactions needed for protein production.
7. Simulate the reactions needed for dimerization.
8. Simulate the reactions needed for protein degradation.

Note that all the stages above are dependent on three primary things:

- Kinetic model and kinetic rate
- Mathematical representation
- Software command line interface

2. Parameters Estimation and model selection

The process of selecting appropriate models, from the pool of tens of possible models, considered as the final aim of biologists or engineers. As indicated before, the biological knowledge, provided by biologists, can be used to rule out some of these models which seems to be implausible. Moreover, researchers can rule out some other models whose behavior do not the match experimental data. Eventually, they will end up with the numbers of candidate models that they cannot rule the out. Each of these models has a number of parameters that determine the

behavior of the model. Generally, model with large number of parameters results in lower errors fitting and they are more likely to over fit the data. On the other hand, these models will produce large errors for any new data sample and they seems to be poorly generalized [4, p.214]. All methods of statistical estimation are performed based on the assumption that the number of parameter to be estimated are fixed and known although their values are unknown. However, in problem of statistical model fitting, the challenge is to appropriately specify the number of unknown parameters [8, p.131].

When reaching the stage, where the engineers have some candidate models, they compare and rank these models based on; model plausibility and model score. Model plausibility can be investigated depending on the biological data and experiments, whereas there are different approaches to assign scores to models such as Bayesian Information Criterion Score (BIC_ score), Akaike's Information Criterion Score (AIC_ score) and Minimum Description Length (MDL_ score) Criterion. Due to limitation in size of this paper, only (BIC_score) is discussed next.

2.1 Bayesian Information Criterion

Based on the knowledge provided by a dataset D, engineers can construct the Bayesian network with structure (S) and estimate a set of conditional probability tables (CPTs). Then Bayesian formula is:

$$P(S \mid D) = \frac{P(D \mid S)\, P(S)}{P(D)}$$

$$\log\ P(S \mid D) = \log\ P(D \mid S) + \log\ P(S) - \log\ P(D)$$

$$\log\ P(D \mid S) \approx \log\ P(D \mid S, CPT) - (K/2) \log\ N$$

Where

D = Observed Data .

N = Number of Data Point in D, Number of Observations, or Sample Size .

K = Number of free parameters to 'be estimated. If estimated model is the linear regression, K is number of regresses.

A formula consists of two terms, the first term measure how well the model, we created, predict the data, and the second term will decrease the score of the model by penalties depending on the number of parameter in model and size of training data [4, p.216]. Furthermore, the form "*-2 log P (D\S)*" and referred as the name (BIC_ score), then from the above can rewrite the formula as:

Note that, the first is on right side, of BIC_score, maximizes the probability function value of estimated model Thus can right formula as:

2.2 Calculating Log-Likelihood value for regression model

BIC_score compare the models to the dataset using the measure of log (Maximum Likelihood Estimate MLE). When simulating, a curve representing the predicted model, will be drawn passing through out all data points. To fit the model to the data it need to minimize the Residual Sum of Squares (RSS) error. That is, get the distance square between data points and model curve as low as possible. If the data set consist the N samples , then:

$$\text{RSS} = \sum_{\text{sample}=1}^{\text{N}} (\text{error per sample})^2$$

When there is more data pints per sample, then:

$$\text{error per sample} = \sum_{\text{data point}=1}^{\text{totaldata point}} (\text{data point} - \text{model fitted value})^2$$

When assuming a single data point per sample , the get:

$$\mathrm{RSS} = \sum_{\mathrm{sample}=1}^{\mathrm{N}} (\text{data poitn} - \text{model fitted value})^2$$

Neglecting assumption of any biases in the model-fitting process , it is assumed that errors of the model per data point is the NII (normal, independent, identical distributed) in statistic. Maximum Likelihood Estimate (MLE) of the model parameters is:

$$\mathrm{MLE} = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^{\mathrm{N}} \mathrm{e}^{-\frac{1}{2}\mathrm{N}}$$

To get this final form of MLE formula should be need know the relationship between RSS and variance ($\mathrm{RSS} = \mathrm{N} * \sigma^2$), and then with some assumptions and process can get the final form of MLE as above. Now from that we can calculate log (MLE) as:

$$\log(\mathrm{MLE}) = -\frac{\mathrm{N}}{2}\log(\sigma^2) \approx -\frac{\mathrm{N}}{2}\log\left(\frac{\mathrm{RSS}}{\mathrm{N}}\right) \approx -\frac{\mathrm{N}}{2}\log(\mathrm{RSS}) + \frac{\mathrm{N}}{2} * \log(\mathrm{N})$$ [4, p221]

3. Conclusion

Many factors are involved in modeling gene regulation starting with the number of cellular species and ending in the environmental factors with some other in between. Adopting a kinetic model, mathematical representation and tools carefully does not ensure the accuracy of model due to the complexity mentioned previously. So, one tries to analyze the observed data of a biological system precisely, construct a asymptotic model and hope for the best when simulating to check the model prediction.

References:

[1] System Biology: definitions and perspective, by Lilia Alberghina, Springer-Verlag Berlin Heidelberg 2005, 2008.
[2] An introduction to system biology, Design principles of Biological Circuit, by Uri Alon, 2007 by Taylor & Francis Group, LLC.
[3] http://rechneronline.de/function-graphs
[4] Computational modeling of gene regulatory network, by Hamid Bolouri, Imperial College Press, 2008.
[5] Daniel T. Gillespie (1977) "Exact Stochastic Simulation of Coupled Chemical Reactions". The Journal of Physical Chemistry 81 (25): 2340–2361
[6] Wikipedia.
[7] stochastic modeling for system biology by Wilkinson, Chapman & Hall/CRC, 2006.
[8] Annals of the Institute of Statistical Mathematics, by Nobuo Inagaki, SprngerLink, 2010