ISBN 978-981-14-4787-7
Proceedings of 2020 the 10th International Workshop on Computer Science and Engineering (WCSE 2020)
Shanghai, 19-21 June, 2020, pp. 359-365
doi: 10.18178/wcse.2020.06.054

Rule-Based Modeling and Simulation of Gene Expression Process

Tianqi Zhao[1], Cuntong Luo[1], Yi Yang[1], Liwei Feng[1] and Xinhua Lu[1,2+]

[1] College of Computer Science and Technology, Jilin University, Changchun, China, 130012

[2] Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University, Changchun, China, 130012

Abstract. A process of gene expression which is designed and modeled by machine language, and by using to process complex data. At the molecular biological level, classifying and sorting the related substances involved in the transcription and translation processes, and using a rule-based modeling method to simulate the gene expression process. By entering the initial data of the designed gene into the program, the computer can give the result of the gene expression process. The experimental data shows that this gene expression result is correct. In this paper, we summarize the rules of the transcription and translation of the gene expression process and the basic part of the central dogma, and then formalize it, and then establish a model to realize the biological information processing in uncomplicated situations, and then realize the gene expression process on the computer.

Keywords: gene expression simulation, rule-based modeling, computer modeling and simulation

1. Introduction

The research of molecular biology has entered a profound and meticulous stage, meanwhile it come to a bottleneck. A chromosome contains gene fragments that may store tens of thousands of base pairs or even more. The network structure of a protein may be very complex. In this case, the traditional experimental way to process and analyse many data gradually becomes unrealistic, and the molecular level experiments are always difficult and cost a long time. With the renewal of modern facilities and the development of computer technology, the research results of molecular biology have been widely used. If the algorithm is designed and the data is given, the computer can process the massive data beyond personal ability [1][2][3].

Gene expression is not only a basic subject in biological research, but also a basic knowledge of molecular biology. The genetic information stored in DNA macromolecules is reflected by the different arrangement order of deoxynucleotides, and the different deoxynucleotides or ribonucleotides are reflected in the different bases. So we can judge different genes by base sequence. However, the number of bases on a DNA is very large, and the analysis of transcription and translation process must be refined to the situation of each base, so it is necessary to introduce a computer to assist in data processing [2].

Rule-based modeling is a modeling method to solve the complexity problem. Instead of manually listing all possible species and reactions that may exist in the system, the rule-based model defines only the reaction motifs within the macromolecular complex and the interactions and modifications of those motifs involved. Rules have the same form as chemical reactions and provide the function of defining different reaction templates by forming specific reaction categories. Through the design of separate and mixed simulation processes of different control mechanisms and the analysis of a large number of experimental data, the effectiveness of rule-based modeling method in reducing the complexity of reaction network system is demonstrated, and the feasibility of extended modeling is also illustrated.

+ Corresponding author. Tel.: +86 0431-85166478.
E-mail address: luxh@jlu.edu.cn, zhongjian5518@mails.jlu.edu.cn.

2. Analysis and Modeling of Transcription Process

2.1. Template Identification

Transcription involves the binding of transcription factors and promoters. It is a process of mRNA synthesis using DNA as template and energy consumption under the catalysis of enzyme [4].

In this paper, the idea of [5] modeling biological network rules is used in the process of gene expression modeling, and the design pattern of biological internal interaction in this literature is followed here.

Definition: for the interaction between biomolecules, it is indicated in the form of reaction rules and indicated by arrow symbols. The arrow is unidirectional. On the left side of the arrow are all kinds of reactants. The part that points to is the product, which can be one kind or many kinds. The common characteristics of a kind of reactants when they interact with a certain substance are called universal properties. For a kind of reactants with general properties, each reaction rule represents an interaction of such reactants [5].

Analyze the process of template recognition, and the reaction rules are as follows:

$$E + F \rightarrow I \qquad [5](1)$$

In the above formula, E strictly represents a kind of action primitives related to all transcription initiation, we only consider promoters here; F represents a kind of catalytic enzymes with regulatory effect. I represent the starting complex. The part to the left of the arrow represents the reactant, and the part to the right represents the product. The meaning of the formula is that the element E(promoter) that represents transcription initiation, and the enzyme f (RNA polymerase) that represents regulation, interact to generate and transcription related complex I.

2.2. Initiation and Extension of Transcription

2.2.1. Transcription of Single Base

For the bases A, C, T and G on the template chain, the U, G, a and C of mRNA are matched respectively. For the transcription of single base, the design formula is as follows:

$$S + B \rightarrow S + T + R \qquad (2)$$

In the formula, S represents the base of the original template chain, B is the substrate involved in the reaction, T is the newly synthesized base, and R is other factors.

2.2.2. Transcription Initiation

Based on the above rule definition, add provisions, and introduce indexes. Index distinguishes different states of a kind of reactant, and the range of index is also the number of reactant states. When the reactant state cannot be determined, the index value uses the wildcard '*' to represent all possible states [5].

Definition: the index can take values in two places within the brackets, such as a(1,2), and the two index values are separated by commas.

Based on single base transcription formula, the following formula groups are obtained by adding index:

$$\begin{aligned} S(1) + B &\rightarrow S(1) + T(3) + R \\ S(2) + B &\rightarrow S(2) + T(4) + R \\ S(3) + B &\rightarrow S(3) + T(1) + R \end{aligned} \qquad (3)$$

2.2.3. Extension of Transcription

Definition: when multiple reactants or products are connected to form a complex, a black dot or percentage sign is often used to indicate the connection of adjacent substances. The left and right order of the two parts of the connection cannot be interchanged.

In the initial process, multiple bases connected after transcription form a mRNA segment, which is called extension process was defined as the mRNA fragment.

After the template recognition is completed, the single character of the expressed base sequence is replaced.

2.3. Splicing of MRNA

The splicing rules of mRNA are as follows:

$$M(1) \cdot M(2) \cdot M(1) \rightarrow M(1) \cdot M(1) + M(2) \qquad (4)$$

M(1) represents exon, and intron M(2) is in the middle of two exons. The position of the product M(1)·M(1) on the right side corresponds to the position of the two M(1) of the reactant successively, that is to say, the M(1) on the left side of the product connector represents the M(1) on the left side of the reactant M(2), and the M(1) on the right side of the connector represents the M(1) on the right side of the reactant M(2). The position cannot be confused.

The head and tail are added to the mRNA after splicing, so the formula of capping is as follows:

$$M(3) + M + M(4) \rightarrow M(3) \cdot M \cdot M(4) \tag{5}$$

In the formula, M(3) represents the added head, M(4) represents the added tail, and M represents the spliced mRNA. The head and tail are respectively connected at the left and right sides of M. So far, the mature mRNA chain has been formed, which carries all the genetic information of the gene and can be transported to the cytoplasmic matrix as the template of translation process to synthesize peptide chain.

2.4. Regulation and Restriction of Transcription Rate

2.4.1. Positive Feedback of Activation Factor

There is a typical positive feedback regulation in the expression of class X genes. Under normal conditions, gene transcription rate is stable, and transcription will form factor R.

Different from the transcription factors of Y genes, X genes generate R(1), which can promote x transcription. Y genes generate R(2), and R(1) and R(2) can be combined as factor R(3), which can inhibit x transcription. The combination of factor and template chain is designed as the combination with base.

Therefore, for individual X gene expression, there is the following formula:

Base transcription generates factor R(1) at a normal rate:

$$S + B \rightarrow S + T + R(1)$$

The resulting R(1) binds to subsequent bases:

$$S + R(1) \rightarrow S \cdot R(1)$$

After binding, the base still does not affect the correct transcription, but this process will speed up:

$$S \cdot R(1) + B \rightarrow S + T + R(1) \tag{6}$$

In this formula group, the binding of S and R(1) does not affect the result of base transcription, only accelerates the transcription rate.

2.4.2. Regulation of Transcription of Two Genes

When X gene and Y gene are expressed at the same time, two factors are generated, and the two factors can also be combined into new substances:

$$\begin{gathered} S + B \rightarrow S + T + R(1) \\ S + B \rightarrow S + T + R(2) \\ R(1) + R(2) \rightarrow R(3) \end{gathered} \tag{7}$$

Combining with the above formula, it can be seen that R(1) will not only combine with the base to accelerate the transcription rate of X, but also combine with R(2) to generate R(3) to inhibit the rate of X and Y. When the transcription rate of X reaches a threshold, the inhibition of R(3) will be strengthened, and the inhibition of R(3) will be weakened when the transcription rate of X decreases to the initial value.

3. Analysis and Modeling of Translation Process

3.1. The Beginning of Translation

The initiation of the translation process also requires the combination of the initiation factor and the structure of mRNA to produce the initiation complex related to translation, and the promoter similar to the transcription process binds to the related substances.

$$K + L \rightarrow V \tag{8}$$

[5]

K represents the initial related structure of mRNA, L represents a variety of related factors, and they combine with poly integrated complex V to start translation.

The starting model of translation is:

$$P(1,3,4) + C \rightarrow P(1,3,4) + Q(12)$$

P(1,3,4) is the starting codon Aug, C is the substrate required for the reaction, and Q(12), that is, m amino acid, is generated.

To realize this process, we need to design the function of identifying the beginning part and the end position. Detect the string representing the base sequence of mRNA, recognize the start and stop sites, translate the string of Aug, generate m, and transfer the data to the next function.

3.2. The Extension of Peptide Chain

When a codon translation is completed, corresponding amino acids will be generated, and each additional amino acid in the polypeptide chain needs to go through carry, transfer, and shift [6].

During the translation process, after the first few links form the peptide chain, the following amino acids are linked to the existing peptide chain.

3.3. The Processing of Peptide Chain

For the translated peptide chain, the types of different segments in the whole peptide chain are marked respectively, and the whole peptide chain is regarded as a complex of different state peptide chains. Design index for three peptide chains as follows:

Table 1: Index of peptide chain types

Type	Head fragment	Functional segment	Tail fragment
Indexes	1	2	3

The processing rules of peptide chain are as follows:

$$W(1) \cdot W(2) \cdot W(3) \rightarrow W(1) + W(2) + W(3) \quad (9)$$

In the formula, W(1), W(2), W(3) are complex connected by peptide chains in three states, that is, the result of internal analysis of the translated W, and the product is the three fragments generated by disconnection. When W(1), W(2) and W(3) are disconnected, W(1) and W(3) are no longer considered, and W(2) is retained as the final result. This is also the result of the whole gene expression process [7].

4. Test Sequence Design

In order to design the test sequence of class X gene, it is necessary to have a complete intron inside the gene, with a start sequence and a stop sequence at the beginning and end.

To design the test sequence of Y class gene, there should be no intron in the gene.

The X·Y gene test sequence was designed to consider the simultaneous expression of the two genes. The start and end base segments of X gene, the start and end segments of Y gene should all be on such test cases.

If a part of DNA contains a promoter, a terminator, and no intron, it is a DNA fragment of y-class gene. Y genes can change the length of their sequences and the sequence of their bases. When the criteria are met, the procedure is expressed no matter how long the sequence is.

If a part of DNA contains two promoters and two terminators, it is a X·Y gene. and Y can be expressed at the same time, and the transcription process will be restricted.

When the input sequence is not the above three types of genes, it is called non-standard gene and cannot be expressed.

For example, the following test cases (X·Y genes)

GGGG AAAA ACTG ATCG ACTG ATCG ACTG ATCG ACTG ATCG ACTG ATCG CACA CACA ACTG ATCG ACTG ATCG ACTG ATCG ACTG ATCG ACTG ATCG TCTC TCTC ACTG ATCG ACTG ATCG ACTG ATCG ACTG ATCG ACTG ATCG GCGC AAAA ACTG ATCG ACTG ATCG ACTG ATCG ACTG ATCG ACTG ATCG GGGG AAAA ACTG ATCG ACTG ATCG ACTG ATCG ACTG ATCG ACTG ATCG GCGC AAAA

Description: This sequence contains two promoters and two terminators. It is a X·Y gene. X and Y can be expressed at the same time, and the transcription process will be restricted.

Ideal mature mRNA:

mature mRNA of X gene is mRNA (1);

mature mRNA of Y gene is mRNA (2).

Ideal final peptide chain:

X gene expression result is peptide chain (1);

Y gene expression result is peptide chain (2).

When the input sequence is not the above three types of genes, it is called non-standard gene and cannot be expressed.

Such as:

GGGG AAAA, ACTG ATCG, ACTG ATCG, ACTG ATCG, ACTG ATCG, ACTG ATCG, only the promoter GGGG AAAA, without the terminator GCGC AAAA, the gene is incomplete and cannot be expressed.

5. Analysis of Experimental Results

After the test sequence is designed, the data document can be established, the program can be run, and the results displayed in the running process of the program can be observed. After the transcription function of the program is completed, mature mRNA data are obtained. After the translation function is processed, the final peptide chain is obtained. After several tests, the correctness and efficiency of transcription and translation functions can be verified by observing and recording mature mRNA and final peptide chain displayed by the program.

The following is the analysis of some experimental results.

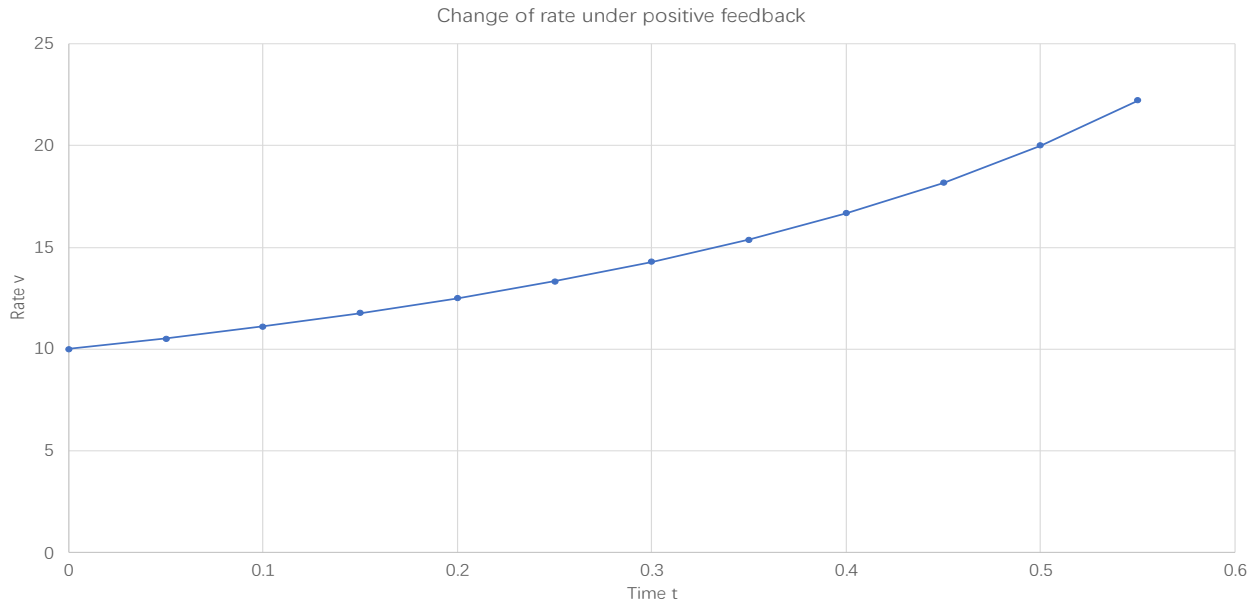

Fig. 1: Change of rate under positive feedback

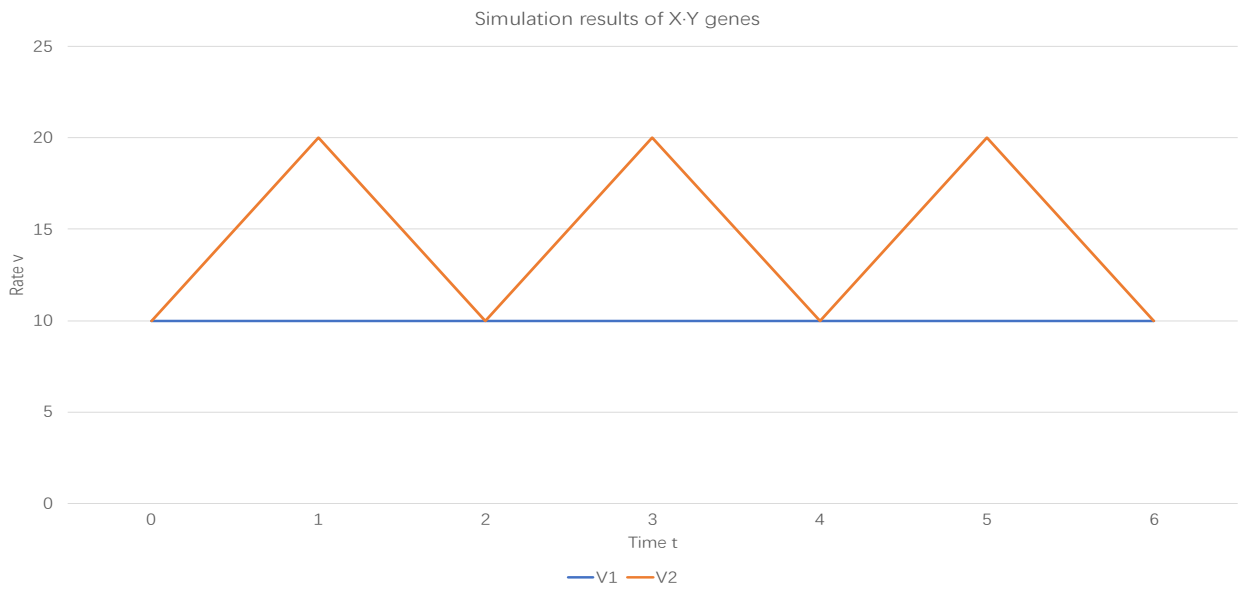

Fig. 2: Simulation results of X·Y genes

Fig. 1 shows the change of rate under the condition of positive feedback of class X gene. The initial transcription rate was set as 10, the change function of transcription rate was 10/(1-kt), and k was the set parameter.

In Fig. 2, the broken line represents the rate of X gene, the flat line represents the rate of Y gene transcription, and the graph reflects the process of X gene transcription rate being restricted. When the initial transcription rate and the rate increasing coefficient of X are set, the rate of X gene transfer increases steadily. When it reaches a certain value, the inhibition of R(3) suddenly increases. The program sets an inhibition parameter to reduce the rate of X gene transfer and keep circulating.

Table 2: Y transcription rate concentration under different setting conditions

Condition	Rate
1; 0	10
1; 0.02	8
1; 0.04	6
0; 0.06	#

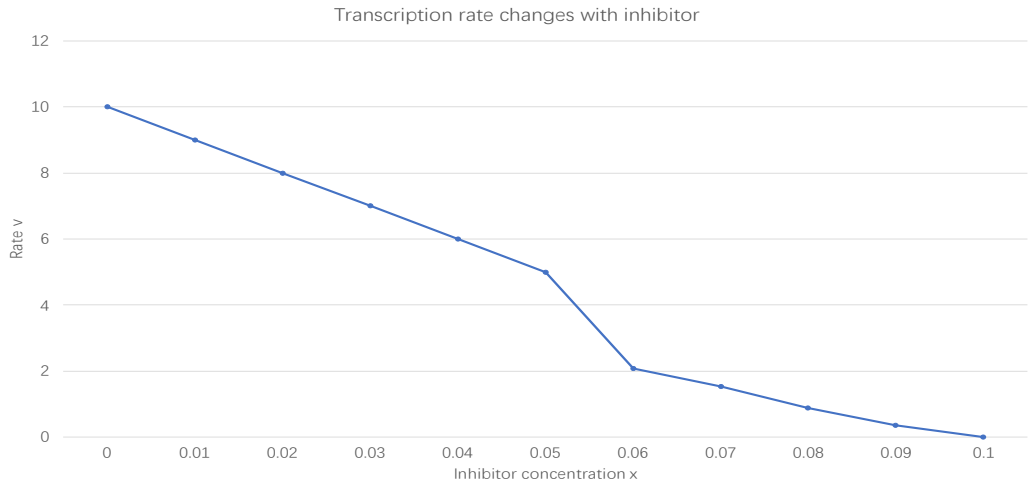

Fig. 3: Transcription rate changes with inhibitor

When the Y gene is transcribed, the program will default to an initial environmental condition: there are various transcriptase and the inhibitor concentration is determined. The result graph of Y gene above shows a transcription rate. To observe the rate of rotation at different inhibitor concentrations, you can modify the initial conditions of the program to display different rates. You can see these in Table 2 and Figure 3

6. Conclusion

In this paper, from the perspective of systematic study of gene transcription and expression mechanism, using rule-based simulation modeling method, a mathematical model for the transcription and translation of class X gene(including an intron), class Y gene(excluding an intron), class X·Y gene is proposed. The simulation results of the model are in good agreement with those of some biological experiments and previous models, which shows that the model is correct.

The success of gene transcription and translation simulation shows that the modeling of cell cycle activity by computer is feasible and effective and has a good prospect of theoretical research and engineering application. In the future work, we can consider improving the model to simulate more gene activities.

7. References

[1] Sean D. Mooney. Continuing challenges swirl around bioinformatics service delivery [J]. Journal of Biomedical Informatics, 2019, 94.

[2] Zhao Yahua. Course of Molecular Biology [M]. Beijing: Science Press, 2011: 155-159,223,295-297.

[3] Ling Zhuqin. Application of molecular biotechnology in the field of environmental engineering microorganism [J]. Engineering Technology, 2016, 12: 2.

[4] Liu Xuejie. MRNA degradation model and kinetic behavior [R]. Guangzhou University , 2017: 74.

[5] Faeder J R, Blinov M L, Goldstein B, et al. Rule-based modeling of biochemical networks. Wiley Subscription Services, Inc. A Wiley Company 2005.

[6] Zhou Chunyan, Yao Libo. Biochemistry and molecular biology [M]. People's Medical Publishing House, 2018: 166-172.

[7] Harris, L.A., Hogg, J.S., Faeder, J.R. Compartmental rule-based modeling of biochemical systems[P]. Simulation Conference (WSC), Proceedings of the 2009 Winter,2009.