Modeling of Protein Production Process by Finite Automata (FA)

ISSAC BARJIS[1], JOE W. YEOL[2], YEONG SOON RYU[3]
Physics and Biomedical Sciences[1], Mechanical Engineering Technology[3]
City University of New York, New York City College of Technology
300 Jay Street, Brooklyn, NY 11201, USA

Department of Mechanical, Aerospace, and Manufacturing Engineering[2]
Polytechnic University
6 Metrotech Center, Brooklyn, NY 11201, USA

Abstract: - The emphasis of this paper is on using Finite Automata (FA) as a modeling tool to model, simulate and analyze the process of protein production. There is tremendous potential for mathematical and computational approaches in leading to fundamental insights and important practical benefits in research on biological systems. Mathematical and computational approaches have long been appreciated in physics and in the last twenty years have played an ever-increasing role in chemistry. Now more and more molecular biologist and biochemists are interested in using computer science applications and mathematical approaches in their work.
In the first section, we introduce the basic concept of bioinformatics and modeling tools in Biology. In the second and third sections the molecular events of protein production process is explained and then based on the molecular events, a Finite Automata model is constructed.

Key-Words: - Protein, DNA, Finite Automata, Modeling, Simulation

1 Introduction

During the latter half of the 20th century, biology was dominated by reductionism approaches that successfully allowed for the generation of information about individual cellular components and their functions. Over the past decade, this process has been greatly accelerated due to the emergence of genomics. We now have entire DNA sequences for a growing number of organisms and are continually defining their gene portfolios.
Also Molecular processes are complex, dynamic and invisible. This complexity makes them difficult to explain, teach, demonstrate and understand. Further more, any laboratory experiment of biological process/reaction is a time consuming business. Many biological experiments require days or weeks, before the dynamic behavior of reaction, process or the expected result can be observed. Even then many biological experiments fail to get the desired or expected result by carrying the experiment once. So the repetitions of the experiments not only make it time consuming business, but also costly business. Pharmaciutical companies are trying to lower the risk and financial burden of clinical trials, so technologies that use computers to model and simulate biological systems are revolutionizing the drug discovery field. Though clinical trials are in no immediate danger of replacement by simulation technologies, pharmaceutical companies are eagerly adopting new means of accelerating the drug discovery and development process.
This study analyzes the emerging markets for technologies that enable researchers to model and simulate biological pathways, cells, tissue, and diseases. It examines upcoming products, pricing strategies, and competitive pressures. With such in-depth coverage, your company will be well prepared to capitalize on a burgeoning industry.
Modeling the structure of Biological molecules and processes are critical for understanding how these structures and processes perform their function. Furthermore, these models can be used for designing compounds to modify or enhance the functions of biological molecules for medical or industrial purpose, or just to simulate a biological process for teaching purpose. Real life laboratory experiments are expensive and time consuming business, and the researchers do not always get the desired results. Using computational approaches to model and simulate a real life laboratory experiment or process would allow us to pre-test the decisions and observe the outcome of the experiment before implementation or real laboratory experiment. This would not only

save time and money, but would also be of a great help for educational institutions to teach the particular process, as the students would observe what actually happens in each step of the reaction throughout the process. These models would allow educational institutions to teach a biological process using modeling and simulation instead of dry text. Analysts could get the outcome of their experiment by touch of few buttons, rather then spending weeks or even months in a laboratory and not always getting the desired/expected results. Computational biology would help researchers to use models of biological processes or experiments, and be more in control of the experiment. Using the model would enable researchers to make any changes in the condition of the experiment or change the precursors of the reaction within the biological process and observe the outcome with matter of minutes instead of spending time and material in the laboratory and not always getting the desired results. Therefore more and more biologist and biochemists become increasingly interested in using computational approaches for their research work, and teachings.

In order to develop a software application for the system under study, one needs to have a proper insight into the essence of action processes of the system. For better understanding of the readers here we very briefly explain the Finite Automata modeling methodology and how it could be applied to the process of protein production.

In this paper efforts of authors from two different disciplines are put together to study the protein production process from computing point of view. Consequently, this paper includes both biological and analytical approaches towards the protein production process.

From the computer science we apply a method to model the protein production process for the mathematical representation of the system. Analysis of the modeling can reveal important information about the structure and dynamic behavior of the system. This information can then be used to evaluate the modeled system and suggest improvements or changes or simply understand the nature of the system under study.

The content of this paper is divided into two sections (a compact model and a detailed model). In section one of this paper, a compact model of the protein production is developed; and in part two, a detailed model of the protein production is presented. Each section consists of two subsections, where the first subsection of each section describes same reactions and molecular events that occur during the process of protein production.

While, in the second subsections models in compact and detailed notations are developed. Conclusions that are derived from the results of this paper are represented in the conclusion part of this paper. Finally, suggestions for future works are given at the end of this paper.

2 A Compact Model of the Protein Production

In this and next sections we will show how model of the protein production process can be built in both compact and detailed notations. Firstly we discuss the protein production process as sequence of DNA, RNA and protein without the molecular events and chemical reactions that occur during this process. Based on this description we will build a model of the protein production in compact notation. A compact model of the protein production covers only core operations (activities) such as transcription, translation, reverse transcription and replication (see figure 2.1). But a detailed model of the protein production, in addition to those operations, covers some molecular events, and precursors/molecules that are required for each step of the protein production process.

2.1 The Protein Production Process

Protein synthesis begins in the cell's nucleus when the gene encoding a protein is copied into RNA. Genes, in the form of DNA, are embedded in the cell's chromosomes. The process of transferring the gene's DNA into RNA is called transcription. Transcription helps to magnify the amount of DNA by creating many copies of RNA that can act as the template for protein synthesis. The RNA copy of the gene is called the mRNA. For the information to be translated from the DNA sequences of the genes into amino acid sequences of proteins, a special class of RNA molecules is used as intermediates [1, 2]. Complementary copies of the genes to be expressed are transcribed from the DNA in the form of messenger RNA (mRNA) molecules. The mRNAs are used by the protein-synthesizing machinery of the cell to make the appropriate proteins. This process, which takes place on sub-cellular particles called ribosome, is referred to as translation. The flow of genetic information in the cell can be summarized by the simple schematic diagram shown in figure 2.1. This figure shows the flow of genetic information. As shown in the figure 2.1, DNA can either replicate to

produce new double helix DNA or can be translated into mRNAs, where these mRNAs would consequently undergo the process of translation to produce secreted protein. Under some special circumstances, mRNAs can undergo the process of reverse transcription and produce double helix DNA.

Figure 2.1 Flow of genetic information

From the above brief description of the protein production process the following conclusions can be made. First, the protein production process is dynamic process, which changes its states after each operation. Second, there some conditional and optional processes take place. For example, DNA can be either replicated or translated into mRNA. Under some special circumstances, mRNA can be, in reverse, transcribed to produce DNA.

Summary

1. The DNA replicates its information in a process that involves many enzymes: Replication
2. The DNA codes for the production of messenger RNA (mRNA) during Translation
3. In eucaryotic cells, the mRNA is processed (essentially by splicing) and migrates from the nucleus to the cytoplasm.
4. Messenger RNA carries coded information to ribosomes. The ribosomes "read" this information and use it for protein synthesis. This process is called translation.

2.2 Finite Automata (FA)

Two major elements in a Finite Automata are states and inputs. All the feasible positions of the pieces on a board call them states. Starting with the initial start state of the pieces on the board can be changed from the state to another based on the value of the number. For each feasible and possible number, only one state meets with the given input number. This continues to get a final state.

A model of this type of rule-based recursive action or machine is usually called a Finite Automata; the term Finite comes from the number of states and the number of inputs being finite; automation is from the deterministic structure or machine (as it is more commonly called), i.e., the change of state is completely governed by the input. Generally, it is called deterministic Finite Automata (DFA) [3-4].

The Finite Automata can be summarized as following:

1. A finite set of states.
2. A finite set of inputs.
3. A finite set of transitions

The state of the machine changes after each instruction is executed, and each state is completely determined by the prior state and the input (thus this is defined as deterministic). The machine simply starts at an initial state, changes from state to state based on the instruction and the prior state, and reaches the final state [5].

2.3 Transition Diagram of the Protein Production in Compact Notation

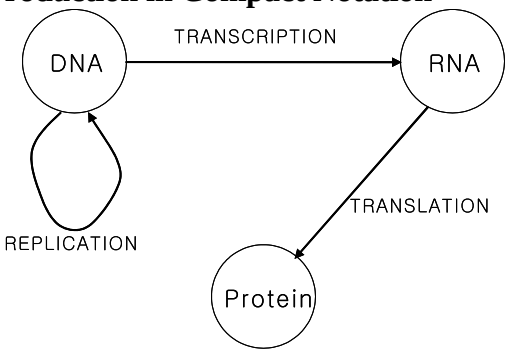

Figure 2.2 *Transition* diagram of DNA to protein via RNA

The transition of genetic information in figure 2.1 is described using a Finite Automata modeling method shown in figure 2.2. The beginning of the process with double helix DNA regarded as an initial product can take two different paths of processes; one is replication, the other is transcription. The process called transcription makes DNA into RNA, which is another product. After that product RNA becomes protein through the process: translation. The arrows in the figure above show that the processes replication, transcription, and translation can not be re-directed or undone, which means that the direction of the state is sequential by a certain order.

However, the results of such models will not be adequate enough and correct, as this model is in a very high level and does not contain enough information of other molecular events and operation, which take place in each steps of this model. In order to develop a detailed and accurate model of protein production process, it is necessary to look at molecular events and reactions that occur in each step (i.e. process) in same details.

		REPLICATION	TRANSCRIPTION	TRANSLATION
START	DNA:	DNA	RNA	·
	RNA:	·	·	PROTEIN
GOAL	PROTEIN:	·	·	·

Table 2.3 *Transition* table of DNA to protein via RNA

Transition table is another representation of a Finite Automata model in figure 2.2. The first column shows the first and final state of the second column.

In the first row, replication, transcription, and translation are the three inputs.

Set of states = {DNA, RNA, PROTEIN}

Inputs = {REPLICATION, TRANSCRIPTION, TRANSLATION}

The rules of Transition to get protein from DNA are following:

state DNA with input REPLICATION, go to state DNA.

state DNA with input TRANSCRIPTION, go to state RNA.

state RNA with input TRANSLATION, go to state PROTEIN

No other rules are applied in this Finite Automata except these three rules.

3 Detailed Modeling of the Protein Production

With reference to a compact model of the protein production it is much easier to go towards further details of the protein production process. In the following two subsections, the protein production process will be described in more detail and the corresponding model will illustrate more detailed information than the previous model (figure 2.2). These details concern some molecular events and precursors/molecules that are necessary for each step of the protein production process.

3.1 The Protein Production Process in Detail

In order to construct an adequate model of the protein production, it is necessary to give same basic information about the molecular events, and precursors/molecules that are required for protein production process in same details. Without these details it is very difficult to construct a detailed and adequate model of the protein production. The genetic information of cells is stored in the form of DNA. This information is used to direct the production/synthesis of RNA molecules or proteins. Each block of DNA that codes for a single RNA or protein is called a gene. Depending on the state of the cell, the genetic information could undergo replication or transcription. If the cell undergoes transcription, it would consequently enter the post-transcriptional processes (tailing and capping, cutting and splicing, transportation), translation, and post-translational process before a polypeptide (protein) can be synthesized [6, 7].

3.2 DNA Replication

DNA exists in the nucleus as a condensed, compact structure. To prepare DNA for replication, a series of proteins aid in the unwinding and separation of the double-stranded DNA molecule. These proteins are required because DNA must be single-stranded before replication can proceed. Chromosomal DNA must be replicated at a rate that will at least keep up with the rate of cell division, and this process is a semi conservative process, i.e. the two strands of the parental DNA duplex act individually as templates for the synthesis of a complementary daughter strand (new strand of DNA) as shown in figure 3.2.In the replication, a parent DNA becomes two identical DNAs; NEW DNA1, NEW DNA2.

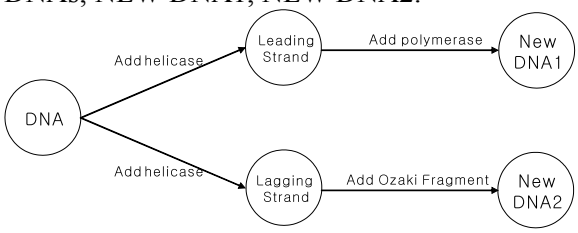

Figure 3.2 Transition diagram of DNA replication

A portion of the double helix is unwound by an enzyme called DNA helicase.

A molecule of a DNA polymerase binds to one strand of the DNA and begins moving along it in the 3' to 5' direction, using it as a template for assembling a leading strand of nucleotides and reforming a double helix.

Because DNA synthesis can only occur 5' to 3', a molecule of a second type of DNA polymerase (epsilon, ε, in eukaryotes) binds to the other template strand as the double helix opens. This molecule must synthesize discontinuous segments of polynucleotides (called Okazaki fragments). Another enzyme, DNA ligase then stitches these together into the lagging strand.

		ADD HELICASE	ADD POLYMERASE	ADD OZAKI FRAGMNET
START	DNA:	LEADING AND LAGGING STRAND	·	·
	LEADING STRAND:	·	NEW DNA1	·
	LAGGING STRND:	·	·	NEW DNA2
GOAL	NEW DNA1:	·	·	·
	NEW DNA2:	·	·	·

Table 3.2 *Transition* table of DNA *replication*

Set of states = {DNA, LEADING STRAND, LAGGING STRAND, NEW DNA1,
NEW DNA2}
Inputs = {ADD HELICASE, ADD POLYMERASE, ADD OZAKI FRAGMENT}
The rules of transition to get new DNA from DNA are following:
state DNA with input ADD HELICASE, go to two states LEADING AND LAGGING STRANDS.
state LEADING STRAND with input ADD POLYMERASE, go to state NEW DNA1.
state LAGGING STRAND with input ADD OZAKI FRAGMNET, go to state NEW DNA2.
No other rules are applied in this Finite Automata except these three rules.

3.3 DNA Transcription to mRNA

From a mechanistic standpoint transcription is quite similar to DNA replication apart from that where in replication only one DNA template strand is transcribed, and only a fraction of DNA strand in a genome is being expressed, and undergoes the process of transcription, in which an RNA molecule complementary to a fraction of DNA strand is synthesized. Transcription begins when DNA dependant RNA polymerase binds to the DNA and moves along the DNA to the transcription unit. Figure 3.3 shows the process from DNA to mRNA.

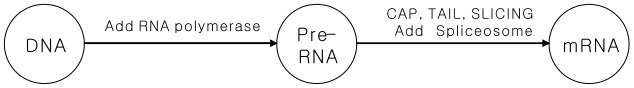

Figure 3.3 *Transition* diagram of DNA *transcription* to mRNA

RNA polymerase binds to DNA to be pre-mRNA product. The product pre-mRNA meets the process called cap and tailing. An enzyme spliceosome is added to the product pre-mRNA which is capped and tailed. In the process of splicing, post pre-mRNA after cap and tail recognize of ribsome. After splicing, mRNA is made. Besides step 1 to 5, DNA is also transcribed into tRNA.

		ADD RNA POLYMERASE	{CAP, TAIL, SLICING, and ADD SPLICEOSOME}
START	DNA:	Pre-RNA	·
	Pre-RNA	·	mRNA
GOAL	mRNA	·	·

Table 3.3 *Transition* table of DNA *transcription* to mRNA

Set of states = {DNA, Pre-RNA, mRNA }
Inputs = {ADD RNA POLYMERASE, {CAP, TAIL, SPLICING, ADD SPLICEOSOME}
The rules of transition to get mRNA from DNA are following:
state DNA with input ADD RNA POLYMERASE, go to state pre-RNA.
state pre-RNA with input CAP, TAIL, and ADD SPLICEOSOME, go to state mRNA.

3.4 mRNA Translation to Protein

Each mRNA codes for the primary amino acid sequence of a protein, using a triplet of nucleotides (called codon) to represent each of the amino acids. In this process mRNA is decoded on ribosome to specify the synthesis of polypeptides (proteins). Following post-transcriptional processing, mRNA transcribed from DNA (gene) in the nucleus, migrates to the cytoplasm (as it is shown in figure 3.4), where mRNAs are read, and proteins assembled, on the ribosome, which are structures composed of rRNA and proteins. Transfer RNA (tRNA) is also needed for translation. Each of these tRNAs can be covalently linked to a specific amino acid, forming an aminocyl tRNA (charged tRNA), and each has a triplet of bases called anti-codon.

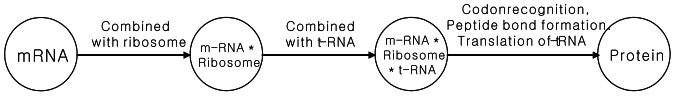

Figure 3.4 *Transition* diagram of mRNA *translation* to protein

After the transcription, mRNA and tRNA are translated to protein through many different processes and with enzymes. In figure 3.4, modeling of translation is described. Ribosome is an already made product where mRNA and Aminoacyl tRNA can be combined to be protein.
The symbol * shows a combined product.
tRNA binds with enzymes aminoacyl tRNA synthetase and amino acid.
After step 1, the product aminoacyl tRNA and mRNA combine in ribosome.

The combined mRNA and tRNA recognized the condon. The process called peptide bond formation is carried out. Translocation of tRNA is processed to be protein. After all the processed, the protein is made.

		COMBINED WITH RIBOSOME	COMBINED WITH tRNA	CODON RECOGNITION, PEPTIDE BOND FORMATION, tRNA TRANSLATION
START	mRNA:	mRNA * RIBOSOME	·	·
	mRNA * RIBOSOME		mRNA * RIBOSOME * tRNA	·
	mRNA * RIBOSOME * tRNA	·	·	PROTEIN
GOAL	PROTEIN	·	·	·

Table 3.4 *Transition* table of mRNA *translation* to protein

Set of states = {mRNA, mRNA*RIBOSOME, mRNA*RIBOSOME*tRNA, PROTEIN}
Inputs = {COMBINED WITH RIBOSOME, COMBINED WITH tRNA, {CODON RECOGNITION, PEPTIDE BOND FORMATION, tRNA TRANSLATION}}
The rules of transition to get protein from mRNA are following:
state mRNA with input COMBINED WITH RIBOSOME, go to state mRNA*RIBOSOME.
state mRNA*RIBOSOME with input COMBINED WITH tRNA, go to state mRNA*RIBOSOME*tRNA.
state mRNA*RIBOSOME*tRNA with input CODON RECOGNITION, PEPTIDE BOND FORMATION, tRNA TRANSLATION, go to state PROTEIN.
No other rules are applied in this Finite Automata except these three rules.

3.5 A Whole Process of Protein Production

The model of whole process from the initial state DNA to the final state protein shows in this section. Through this modeling process, the principal products, enzymes, process and data flows represent.

Set of states = {DNA, LEADING STRAND, LAGGING STRAND, NEW DNA1,
NEW DNA2, Pre-RNA, mRNA, mRNA*RIBOSOME, mRNA*RIBOSOME*tRNA, PROTEIN}
Inputs = {ADD HELICASE, ADD POLYMERASE, ADD OZAKI FRAGMENT, ADD RNA POLYMERASE, {CAP, TAIL, SPLICING, ADD SPLICEOSOME, COMBINED WITH RIBOSOME, COMBINED WITH tRNA, {CODON RECOGNITION, PEPTIDE BOND FORMATION, tRNA TRANSLATION}}
The rules of transition to get the final state protein from the initial state DNA are following:
state DNA with input ADD HELICASE, go to two states LEADING AND LAGGING STRANDS.
state LEADING STRAND with input ADD POLYMERASE, go to state NEW DNA1.
state LAGGING STRAND with input ADD OZAKI FRAGMNET, go
to state NEW DNA2.
state DNA with input ADD RNA POLYMERASE, go to state pre-RNA.
state pre-RNA with input CAP, TAIL, and ADD SPLICEOSOME, go to state mRNA.
state mRNA with input COMBINED WITH RIBOSOME, go to state
mRNA*RIBOSOME.
state mRNA*RIBOSOME with input COMBINED WITH t-RNA, go to state mRNA*RIBOSOME*tRNA.
state mRNA*RIBOSOME*tRNA with input CODON RECOGNITION, PEPTIDE BOND FORMATION, tRNA TRANSLATION, go to state PROTEIN.
No other rules are applied in this Finite Automata except these eight rules.

4 Conclusion

The mean purpose of this paper was to develop modeling methodology for the production of proteins. This methodology helps formalization, modeling and simulation of the production of proteins. Therefore the first conclusion is that dynamic processes of molecular and biological systems in general, the protein production process in particular can be modeled as a discrete dynamic system. Two areas can benefit from such a methodology that has been presented in this paper: to stimulate research and to assist teaching. For the teaching purposes, this can assist to visualize the protein production processes model from state to state and to explain how all molecular events, reactions and operations together provide production of proteins from DNA. It can show how the precursors and substrates, which are required for each step of the protein production processes, are bound to their targets. This paper can be also useful for the training program offering molecular biology with modeling and information sciences integrated into the individual courses, to train students in the use of computational techniques in the study of molecular and biological science.

For the research purposes, one can use this methodology for the protein production modeling and simulation. It is also useful for protein and DNA sequence analysis. Finally, it seems that the results of this paper are one of the first efforts to apply discrete systems modeling technique to molecular-biology processes. In its turn it is another one step towards bringing computer science and molecular biology closer and calling it bioinformatics.

References:

[1] Stryer,L., 1995. Biochemistry, (4th edition). Freeman, USA.

[2] Hawkins,J.D., 1997. Gene Structure and Expression. Cambridge University Press.

[3] http://www.cs.brown.edu/research/ai/dynamics/

[4] Baxevanis,A.D. and Ouellette,B.F.F., 1998. Bioinformatics - A Practical Guide to the Analysis of Gene and Proteins.

[5] http://www.cs.tau.ac.il/~bchor/CM/

[6] Alberts,B., Bray,D., Lewis,J., Raff,M., Roberts,K., and Watson,J.D., 1994. Molecular Biology of the Cell, (3rd edition). Garland.

[7] Karp,G. 1996. Cell and Molecular Biology. Wiley.

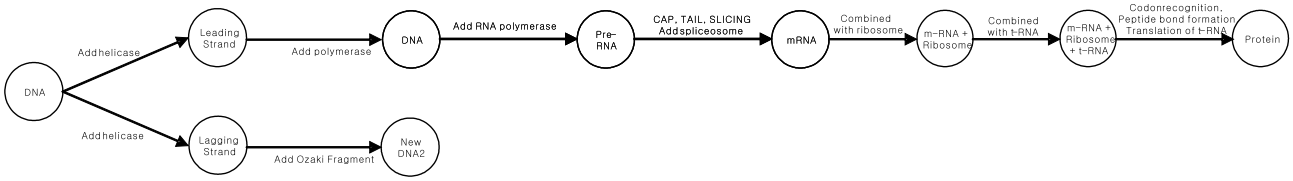

Figure 3.5 ***Transition*** **diagrams of DNA** ***replication, transcription, and translation***

		ADD HELICASE	ADD POLY MERASE	ADD OZAKI FRAGMNET	ADD RNA POLY-MERASE	CAP, TAIL, SLICING, and ADD SPLICEOSOME	COMBINED WITH RIBOSOME	COMBINED WITH t-RNA	CODON RECOGNITION, PEPTIDE BOND FORMATION, t-RNA TRANSLATION
START	DNA:	LEADING AND LAGGING STRAND	·	·	·	·	·	·	·
	LEADING STRAND:	·	NEW DNA1	·	·	·	·	·	·
	LAGGING STRND:	·	·	NEW DNA2	·	·	·	·	·
	NEW DNA:	·	·	·	Pre-RNA	·	·	·	·
	Pre-RNA:	·	·	·	·	mRNA	·	·	·
	mRNA:	·	·	·	·	·	mRNA * RIBOSOME	·	·
	mRNA * RIBOSOME:	·	·	·	·	·	·	m-RNA * RIBOSOME * t-RNA	
	m-RNA * RIBOSOME * t-RNA:	·	·	·	·	·	·	·	PROTEIN
GOAL	PROTEIN:	·	·	·	·	·	·	·	·

Table 3.5 ***Transition*** **table of DNA** ***replication, transcription, and translation***