Modeling genetic switch of Epstein-Barr virus

-Cooperativity, competition & switch sensitivity

LIZHE ZHU

Department of Microtechnology & Nanoscience
CHALMERS UNIVERSITY OF TECHNOLOGY
Göteborg, Sweden 2007

THESIS FOR THE NANOSCALE SCIENCE AND TECHNOLOGY
MASTER'S PROGRAM

Modeling genetic switch of Epstein-Barr virus

Cooperativity, competition & switch sensitivity

Lizhe Zhu

CHALMERS

Department of Microtechnology and Nanoscience
CHALMERS UNIVERSITY OF TECHNOLOGY
Göteborg, Sweden 2007

Modeling genetic switch of Epstein-Barr virus

-Cooperativity, competition & switch sensitivity

LIZHE ZHU
lizhe@student.chalmers.se

Work completed at

Computational Biological Physics
KTH-Royal Institute of Technology
SE-100 44 Stockholm
Sweden, 2007

Abstract

Epstein-Barr virus (EBV) is a widely spread human lymphoma virus that conducts its persistent infection when we are in childhood or adolescence. Each year there is a huge number of tumors associated with this virus. In the present work, I am investigating the gene regulation on a multiple binding sites region of its genome FR, where an encoded protein and a human protein compete to bind and determine the behavior of the genetic switch of the C promoter. Attention is given to the influence of the cooperative and competitive binding effects of the two proteins on the sensitivity of the genetic switch and how cooperativity and competition affect each other.

Keywords: Epstein-Barr virus, genetic switch, switch sensitivity, cooperative binding.

Contents

List of figures

List of tables

Glossary

Lymphoma Cancer of the lymph nodes and spleen that causes excessive production of lymphocytes

Lymphoblasts Dividing lymphocytes

Chloroplast The organelle that carries out photosynthesis and starch grain formation.

Mitochondrion A eukaryotic organelle that is the site of ATP synthesis and of the citric acid cycle. The eukaryotic cellular organelle in which the Krebs cycle and electron transport reactions take place.

Nucleoid A DNA mass within a chloroplast or mitochondrion.

Herpes virus One of a group of DNA containing viruses causing latent infections in man and animals

1 Introduction

Epstein-Barr virus, discovered from examining electron micro-graphs of cells cultured from a common childhood tumor in sub-Saharan Afirca--Burkitt's lymphoma in 1964, is now known to correlate with this and other types of cancer. Being latent after its infection into a child or a teenager, this virus disturbs the cell cycle and forces the normal human B-cells (a species of white blood cells) to proliferate, which is possibly linked to formation of tumors.

Compared with higher living systems, viruses are relatively simple to study. With help of tools from molecular biology, genes and proteins involved in the maintenance and mechanism of virus' transition between different latency states are gradually being classified.

1.1 Goals

Transition between the latency and lytic state, and among different latency states are critical behaviors of Epstein-Barr virus. Previous work by the group has provided a good framework of statistical model [18]. The core part of this effort is considering of competitive binding to the FR region of the viral DNA between two proteins Oct-2 and EBNA-1. While Oct-2 is a human protein, EBNA-1 is self-contained in the viral genome and essential for maintenance of the viral genome, and the transition of viral survival strategy.

The present work is based on the previous statistical model, aiming at adding cooperative effects of EBNA-1 and Oct-2 to the model so that cooperativity and competition can both be considered. Focus is given to their influence on the sensitivity of the genetic switch of C promoter and their influence on each other.

1.2 Outline

Following the introduction, necessary background of molecular biology is described

from a physicist's point of view. Thereafter, some statistical mechanics as base of the model is introduced and attention is paid mainly to describing genetic regulation. Chapter 4 gives a short description of the study of genetic networks and serves to provide broader background for this work. Biological knowledge of Epstein-Barr virus is described in Chapter 5, with focus on working mechanism of C promoter. Chapter 6 is devoted to basic concepts of combinatorics and several paradigm problems used in the core part of algorithm in this work. Chapter 7 describes how algorithms are developed for different cases of problems. Major results of the present work are given as Chapter 8 while discussion forms the final chapter.

2 Molecular biology

This section introduces basic knowledge of molecular biology necessary for this thesis. Central dogma is introduced first, followed by more detailed description of general molecular mechanisms of gene transcription both in prokaryotes and eukaryotic cells.

2.1 Central dogma of life

Most activities of the cell, the minimum unit of any life form, are controlled by two main classes of molecules namely proteins and nucleic acids. Among them deoxyribonucleic acid (DNA) is the most basic and the term "gene" is usually defined as a piece of DNA molecules with active inheritance meaning. One hypothesis is that the only goal of life is to preserve the "selfish" genes under the pressure of natural selection including competition among the genes themselves [29]. Though in some life forms, the genetic information is coded on RNA molecules, we here grant for convenience that all activities of an organism, unicellular or multi-cellular, are ultimately controlled by genes defined on DNA only.

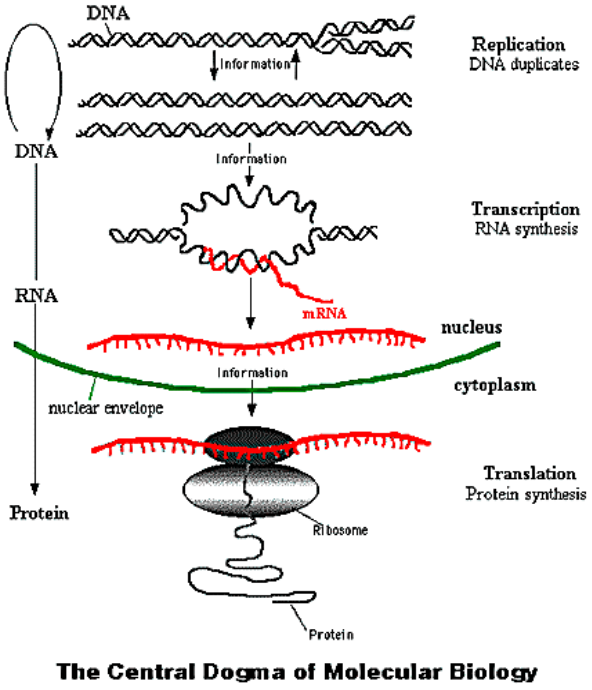

Figure 2.1 Central dogma [30]

DNA preserves the ultimate living information via self-replication and plays as information source of living processes by passing the information to messenger-RNA (mRNA) through transcription. And these mRNA molecules, of relatively smaller size, travel in cytoplasm and reach their destination a type of organelle called ribosome, where translation happens. Coded information is finally transferred from mRNA to ribosomes, which read the codes and use it for protein synthesis. All these processes are involved in almost all biological activities, structural or enzymatic and are summarized as Central Dogma of molecular biology. The central inheritance role of nucleic acids is explained not only by the procedure of the central dogma but also by the fact that proteins contain no information for production of proteins, DNAs or RNAs.

Except in inheritance, proteins rather than nucleic acids are the dominant living polymers in both number and variety in the cell. Structurally, proteins constitute an integral part of the cell membranes, pigments, ion channels, etc. Functionally, in the form of enzymes, proteins play vital roles in metabolism processes, catalyzing various reactions in the cell. It might be appropriate to compare the relationship between the nucleic acids and the proteins to the one between the commander and the soldiers, where DNA gives order to RNA and RNA transfers the order further to functional proteins for actions, while DNA's replication and other survival activities are reversely served by the proteins. Thus, uncovering secrets of life processes seems to be searching for the mapping from the space of the inheritance material (genotype) towards the function space (phenotype).

2.2 Transcription mechanism & genetic regulation

Expression of genes is regulated by some proteins—themselves products of genes. The control mechanism ensures that the cell is not unnecessarily flooded by the proteins, enzymes for enormous number of reaction, required only at particular times and only in a particular quantity. In this section, we will consider the general molecular mechanisms of the first step of protein synthesis in the Central Dogma—transcription, as it is our major research object later for the Epstein-Barr virus.

Transcription is a process where mRNA is formed on basis of the DNA molecule with help of specific enzymes called RNA polymerases (RNAp). Forms of such enzymes differ between prokaryotic cells and eukaryotic cells.

In prokaryotes

RNAp of prokaryotes consists of two parts- a core enzyme and a sigma factor. The sigma factor initiates transcription of mRNA by recognizing and binding to some pieces of DNA (promoter) with characteristic genetic codes. The core enzyme then travels along the DNA template, reads the following codes and synthesizes mRNA continuing transcription. But sometimes RNAp will neither bind to the promoter nor continue transcription easily. At some promoters, RNAp can be blocked by proteins or other types of living polymers bound to some sites of DNA on the way (operator) when trying to travel along the template. These proteins inducing blocking of RNAp are named repressors. In bacteria, RNAp's binding to DNA can be enhanced by some other proteins—inducers. But the enhancement is not always necessary-transcription at many promoters can be activated without help from inducers as RNAp binds directly to the gene. Thus, expression of genes in prokaryotes is often at high rate unless turned off by a repressor [25].

In eukaryotes

In eukaryotes, however, the enhancement becomes a pre-condition. Eukaryotes contain three types of RNA polymerase. Among them, RNAp II, which contains approximately ten subunits and transcribes a wide array of genes, can not correctly initiate transcription alone. In fact, it binds not directly to the genes, but rather to one or more types of proteins called transcription factors. Some these factors are site-specific, recognizing characteristic sites on DNA and decide where for RNAp to bind, while some others are general transcriptional factors that do not have the site-specific feature. Existence of these factors explains well the observation that eukaryotic genes are often expressed only at very low levels or not at all—transcription can not happen unless turned on by one or more transcriptional activators. The mechanism also complicates

the transcriptional patterns, as expression one gene is then controlled by multiple types of factors, which are products of other genes.

3 Statistical physics in genetic regulation

Statistical physics is one of the fundamental theories of physics, and uses methods of statistics to solve physical problems. It can describe a wide variety of fields with an inherently stochastic nature. Also it plays a role of bridge between the microscopic and macroscopic world, defining properties of a system as an assembly of many similar small entities.

3.1 Key concepts & equations

The most key elementary quantity in statistical mechanics is the entropy S. Here we start with Boltzmann's microscopic definition of entropy. Suppose the number of possible microstates corresponding to the macroscopic state of a system of an isolated system is given by Ω. Entropy is then defined as:

$$S = k_B \ln(\Omega)$$

,where k_B stands for the Boltzmann Constant ($8.617342 \square 10^{-23}$eV/K). Thus the probability of finding this system in some macroscopic state with entropy S is proportional to e^{S/k_B}.

Now we consider conceptually a simple system Θ isolated from its surroundings. It contains two subsystems. We concentrate on one of the subsystems, the small System A and consider the other one, System B as surrounding reservoir. We characterize the system here by only two extensive state variables, the energy E and the entropy S. Thus we have the following equations:

$$E_\Theta = E_A + E_B = const.$$
$$S_\Theta(E_\Theta) = S_A(E_A) + S_B(E_B)$$

Following the first law of thermodynamics, the total energy of Θ is fixed. Recall the Fundamental Equation of thermodynamics:

$$\frac{dS}{dE} = \frac{1}{T}$$

where T is temperature. When in thermal equilibrium, System A and B approach the same temperature T^0. Assume that A and B are in equilibrium and that the reservoir B is large enough so that the overall temperature is fixed at $T^0 = \frac{dE_B}{dS_B}$. We give a slight perturbation to energy of the whole system, say $\Delta E_A = E_A - E_A^0$, and this generates a change in energy of B, $\Delta E_B = -\Delta E_A$.

Entropy of Θ when A has the energy of E_A is then,

$$\begin{aligned} S_\Theta(E_A) &= S_A(E_A) + S_B(E_B^0 + \Delta E_B) \\ &= S_A(E_A) + S_B(E_B^0) + \frac{dS_B}{dE_B}\Big|^0 \Delta E_B \\ &= S_A(E_A) + S_B(E_B^0) + \frac{1}{T^0}\Delta E_B \\ &= S_B(E_B^0) + S_A(E_A) - \frac{1}{T^0}\Delta E_A \end{aligned}$$

Thus, $P(E_A)$ denoting the thermodynamic probability of system A in energy E_A becomes

$$\begin{aligned} P(E_A) &\propto e^{\frac{S_\Theta(E_A)}{k_B}} \\ &\propto e^{(S_B(E_B^0) + S_A(E_A) - \frac{1}{T^0}\Delta E_A)\frac{1}{k_B}} \\ &\propto e^{(S_A(E_A) - \frac{1}{T^0}E_A + \frac{1}{T^0}E_A^0)\frac{1}{k_B}} \\ &\propto e^{\frac{S_A(E_A)}{k_B}} e^{-\frac{E_A}{k_B T^0}} \end{aligned}$$

While the first term counts the degeneracy of states of A with energy E_A, the second term counts the probability that one state at energy E_A is selected. Consider two energy states of A, E_A^1 and E_A^2. We can get the famous Boltzmann weight factor:

$$\frac{P(E_A^1)}{P(E_A^2)} = e^{\frac{-(E_A^1 - E_A^2)}{k_B T^0}} = e^{\frac{-\Delta E_A}{k_B T^0}}$$

It states that the ratio of the probability of a system with two energy levels is given by $e^{\frac{-\Delta E}{k_B T^0}}$. We will use Boltzmann weight heavily in our model.

3.2 Representing genetic regulation

The complex process of transcription can be explained as binding among various types of molecules, especially those between the regulatory proteins and the DNA molecule. Here we find statistical physics efficient describing whether a molecule P(protein) is bound or not bound to another molecule D (DNA). Bound or not bound are two states of the system Θ, made up by molecule A and B. Thus we can measure the energy of each of these two states E(A bound), E(A free) and then use the Boltzmann weights to get the relationship of the thermodynamic probability of them:

$$\begin{aligned} P(E_{bound}) &= e^{\frac{-(E_{bound} - E_{free})}{k_B T}} P(E_{free}) \\ &= e^{\frac{-\Delta E}{k_B T}} P(E_{free}) \\ &\propto e^{\frac{-\Delta E}{k_B T}} [A_{free}] \end{aligned}$$

By measuring concentration of the free molecule A in a cell, we can then know how many As have bound to a single site. Summing over all energy levels we can get the canonical partition function $Z_{canon} = \sum_i [A_{free}^i] e^{\frac{-\Delta E_A^i}{k_B T}}$. The probability of A bound is then $P = \frac{1}{Z_{canon}} [A_{free}] e^{-\frac{E_{free}}{k_B T}}$.

Applying this to all regulatory proteins generates then a series of complex patterns, each in accordance with a certain level of the gene expression [5]. Using grand canonical ensemble, we can treat a DNA sequence consisting of many binding sites. Grand canonical partition function is obtained by summing over all possible numbers of binding sites:

$$Z_{grand} = \sum_{n=0}^{N} Z_{canon}(n)$$

Then, the probability of n As occupying the region is

$$P = \frac{1}{Z_{grand}} [A]^n e^{n\frac{-E_A}{k_B T}}$$

4 Gene regulatory networks

On the way of understanding life processes, an influential argument is the analogy between the cells and a complex machine. This might recall us of what we know about the artificial electrical machines, which have been studied for decades, primarily thanks to development of control theory. In designing an electrical system, one of the most essential works is to build functional building blocks, with help of our knowledge of the electrical loops. Are these concepts from control theory applicable to living systems? Intuition drives us to believe it, but it is just within recent decades can we be confident at its usefulness.

Confidence then comes from two directions. On one hand, molecular biology enables us to look at molecular details on relatively simple organisms, e.g. virus, bacteria. Take viruses for example, interactions and functions of the core inheritance material DNA and the proteins that regulate the expression of some of them, have been investigated almost thoroughly via molecular biology techniques. On the other hand, physicists have proved modeling of genetic networks to be efficient tools for explanation of biochemical experimental data. Such study of large scale genetic networks finds good opportunity of applying what we know already in control theory. This fact popularizes "systems biology" as a joint name of the study of genetic or protein networks.

4.1 Genetic networks

From the informatic perspective, we can describe working mechanism of life as a flow of information. Due to our better understanding of the specificity of information processing in living systems, compared with chemical or physical systems, such way of thinking has become increasingly essential in life science. Figure 4.1 gives a broader illustration concerning information flows among all living molecules.

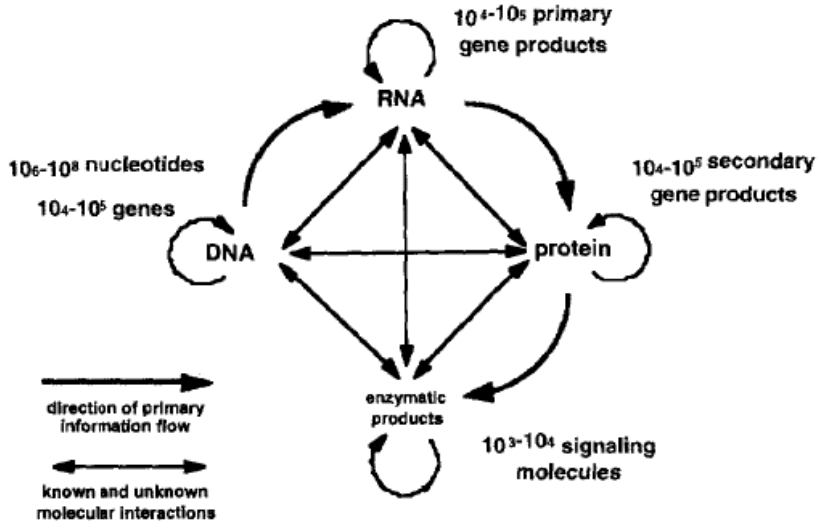

Figure 4.1 Biomolecular information flow and feedback regulation

As mentioned previously, genetic control on protein synthesis is regulated by some of its own products. This means the information flow is never in one direction but forms very complex networks, including feedback. Despite the primary information flow –the central dogma, information can also spread through other interactions among living molecules. For example, certain RNA can be enzymes catalyzing its own replication reaction; RNA interference to expression of DNA has also been proved [21].

4.2 Less is more

Theoreticians have built many interesting models of complex networks trying to explain life process on larger scale, taking into account enormous varied interactions to see "emergence" of new properties not shown in parts of the systems. This field is usually referred to as complex systems.

Practically, biochemical data is explained by different scales of models. Major procedure of such attempt can be abstracted as two steps: 1) Investigate properties and functions of conceptual networks and categorize them 2) Analyze biological data and try to decide to which group of networks the genetic network belongs.

However, data is never easy to acquire, especially for higher organisms with large scale complex genetic networks. Take our modeling on Epstein-Barr virus (EBV) for

example, it is based on over 40 years' investigation by molecular biologists. It is then not unimaginable that living systems of a bit more complexity can be far more time-consuming than the virus via current experimental techniques. Compared with ambitions of physicists and chemists longing for quantifying complex biology, such slowness is like "a mountain too high to climb", blocking efficient acquisition of data.

As a result, we should still try to find efficient way to reduce the gross of data as much as possible. One candidate method lies in reducing the number of transcription levels. Expression level of genes is usually described by biochemical kinetics, series of differential equations, where it is changing continuously. If we can discretely describe the expression levels, each in accordance with one or several biological meanings, many biochemical details can be then neglected and the model can be simplified, telling more stories.

Is this simplification possible? Proofs exist in eukaryote, but mainly in bacteria. Our work can be one of the examples in eukaryote, where two states are enough to describe one of the critical behaviors of Epstein-Barr virus. This is usually referred to as genetic switch. Validity of such simplification will also be given in later chapters by the sharpness of the probability of the switch to be "ON". Genetic networks in which each gene holding only two states are referred to as boolean networks. They have been proved useful in systems biology, since switch-like dynamics is found in other viruses and many other organisms [3].

4.3 The parts or the whole?

Progress in physics and chemistry last century was dominated mainly by reductionism, so was that of biology. Breaking life down to physical or chemical problems proves to be effective and has made great contributions to the development of many applications, such as pharmacy and clinic science. The role of efforts in conventional fashion can not be replaced, though it is now an age of systems biology. It is encouraging that we can study large scale biological networks, but detailed chemical and physical investigation of the molecular mechanism is still of great importance.

5 Epstein-Barr virus

Nature creates and cultures enormous categories of living components in an incredibly efficient way, i.e. as expressed in informatics terms, nature holds "unlimited amount of information". However, our life is finite. In greater detail we want to know about them through experiments, the more difficult it is to carry these experiments, including studies of Epstein-Barr virus. This piece of work is based on over 40 years' biological research of EBV.

5.1 Glimpse into virology

Virology witnessed its beginning back in the end of 19th century, when a young Russian scientist, Dimitri Ivanovsky presented his famous work on "filterable agent" that causes tobacco mosaic disease. Ivanovsky's age was a time in which spontaneous generation of variation of organism ruled the understanding of diseases. The new definition of causality of disease from a "filterable agent" opened up a window to scientists to build up outline for today's research of infectious disease.

5.2 History of EBV

The study of Epstein-Barr virus can be traced back to the electron microscopy examination of Burkitt's malignant lymphoma, done by Epstein M.A., Barr Y.M. and colleagues over 40 years ago. The original purpose of the investigation was to suggest a causation of a transmissible vector-borne agent by considering the climatic and geographical factors of this lymphoma.

Cells were first cultured stationarily in two separate environments for 75 and 82 days respectively and then collected in suspension by drawing the culture fluid so that they could grow free-floating. After a series of chemical treatment and operation of slicing, cells were examined under Philips EM 200 electron microscope. The general structural organization of the lymphoblasts was remarkably uniform and not difficult to recognize. However, among a small number of the cells, unusual structure, the

morphology of which indicated a virus, was observed both in the cytoplasm and the nucleus. These observed virus particles were at different stages of their maturing. While the immature particles, some with a nucleoid -inheritance part of the virus- and some without, were found across the cells, the mature ones were only found in the cytoplasm, indicating that assembly of these particles happened between the cellular membrane and of the nucleus.

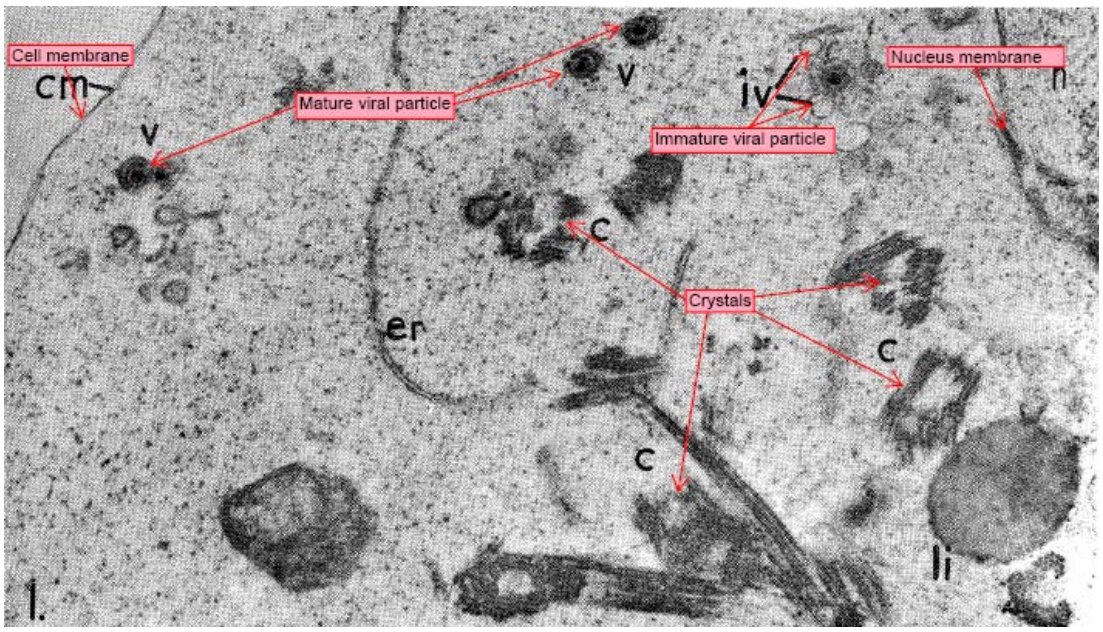

Figure 5.1 First discovery of EBV [22]

Not surprisingly, this virus was named after their finders as EBV, as it is called now. Interest was then directed towards the mechanism of the maturation process of the virus in the lymphoblasts and its relation to other herpes virus in the tissue culture. One of the most important features of the virus known from this first discovery is that "a 'passenger' role can be assumed since the agent has persisted in vitro in the dividing cells for many weeks." That is the constant latency this virus show after the original infection of the host cells.

Later it was found that distribution of such gamma-herpesvirus was far from restricted but widely spread in all human populations and the latent infection as described "several weeks" in the original discovery report is actually life-long in the vast majority of individuals[23].

5.3 Genetic switch in EBV

Infection of EBV happens usually at our childhood or adolescent time. This virus spreads widely over 90% of population of the world. Each year, there are 170,000 tumors associated with it.

Everything begins with a kiss

Through kissing from an infected adult to a child, this virus penetrates successfully the kid's epithelial cell and enters its immune systems following circulation of its blood and lymph. Though detailed mechanism of this primary infection in vivo is not fully established yet, we here try to present a rough description.

Figure 5.2 Infection of EBV

In the blood and lymph, B-cell is one of the major types of lymphocyte performing the role of immune surveillance. Each B cell has a unique receptor protein on its surface that will bind to one particular antigen. Using this receptor, each B cell is able recognize a specific type of antigen, the threat to the organism. Before exposed to various antigens (receptor not taken for once), a B cell is called Naive B cell. After binding of the antigen to the receptor of the Naive B cell, it may be differentiated into different types of cell with respect to different function it serves thereafter.

Naive B cell is believed to be the primary target of the EBV able to invade the lymphoid tissues. Caught by the receptor of B cells, EBV injects its genome and begins its own life cycle in the host cell.

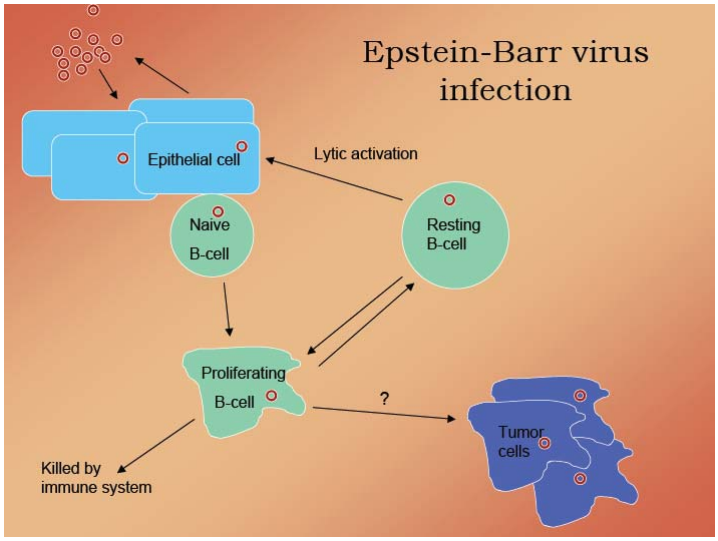

Figure 5.3 Survival strategy of EBV [Courtesy of Maria Werner]

Figure 5.3 shows the important cell programs of B cells directed by EBV. After infection, some host B cells are driven by the viral genes to proliferate. This program of proliferating is referred to as Latency III, the latent growth-transforming infection. Many of these proliferating cells are recognized and removed by the immune system. But some of them manage to escape by suppressing expression of the viral genes and establishing a stable reservoir of resting viral-genome-positive memory B cells (Latency I), which later can infect the epithelium again and function as a source of persistent infection. Some resting memory B cells can be reactivated and switch back to Latency III under immune suppression. It has been suggested the rapid proliferation of some of these reactivated B cells is strongly associated with tumor formation.

Molecular Basis

Within latently infected B cells, the viral genome is maintained as a circular plasmid of approximately 165,000 base pairs, as shown schematically in Figure 5.4. In resting B cells, this plasmid replicates semi-conservatively once each cell cycle and is then partitioned faithfully as the host cell genome to daughter cells. OriP denotes the region where replication of the viral genome happens.

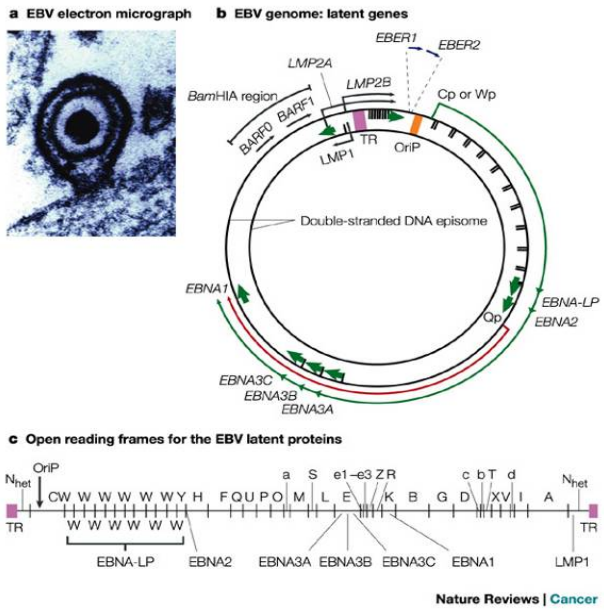

Figure 5.4 Genome of EBV [23]

ENBA-1

Replication of the viral genome is auto-regulated by EBNA-1, a protein both encoded in the viral genome and regulating expressions of the viral genes [24]. This can also be noticed from Table 5.1, as EBNA-1 is the only encoded protein expressed in all latency programs of the virus. EBNA-1 behaves as the only bridge among all programs (Latency I, II, III) after primary infection; the concentration of EBNA-1 in the host cell determines which program the cell shall be directed to.

Program	Genes Expressed
Latency 0	EBER-1&2, LMP-2a
Latency I	EBER-1&2, LMP-2a&b, **EBNA-1**
Latency II	EBER-1&2, LMP-1,LMP-2a&b, **EBNA-1**
Latency III	EBER-1&2, LMP-1,LMP-2a&b, **EBNA-1**~6

Table 5.1 Proteins produced from the EBV genome.

C promoter & Oct-2

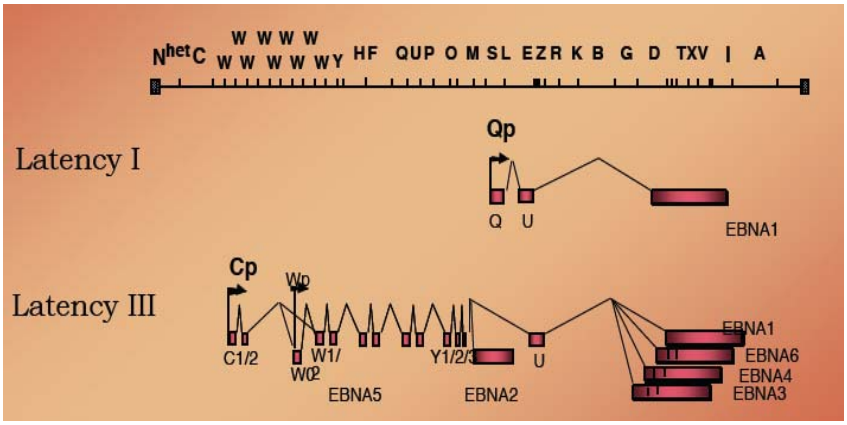

Figure 5.5 Latency I & III associated with Cp and Qp [Courtesy of Maria Werner]

Latency I and III are associated with two promoters Cp and Qp, respectively. At molecular level, Latency I equals Cp OFF but Qp ON, while Latency III means Cp ON but Qp OFF.

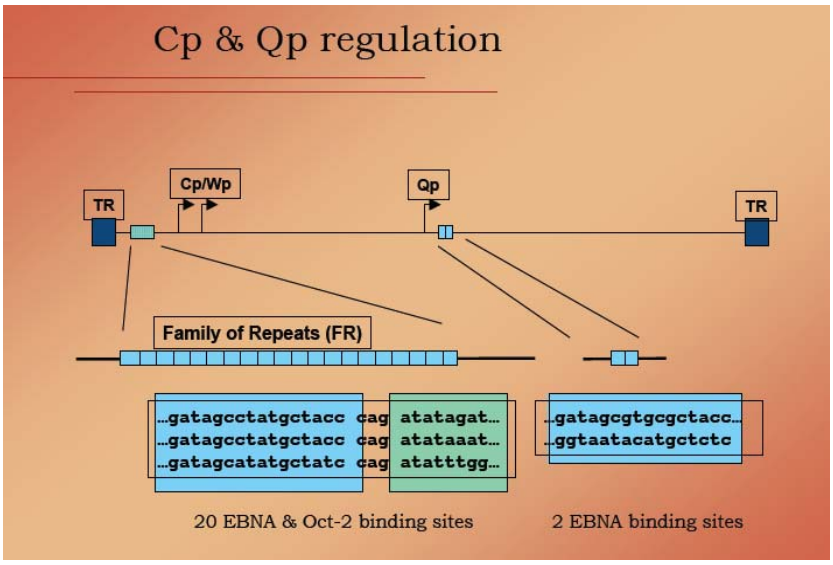

Figure 5.6 Cp & Qp regulation [Courtesy of Maria Werner]

Qp is regulated by a region on its right, which consists of two binding sites of EBNA-1. If any of these two sites are bound with EBNA-1, RNAp will be blocked so that, all genes on the right of Qp can not be expressed, i.e. Qp will be OFF.

Activity of Cp is determined by a 40-site region, Family of Repeats (FR) on its left. Throughout FR, 20 binding sites of EBNA-1 and the rest 20 of Oct-2, a human protein, lie side by side.

Family of Repeats

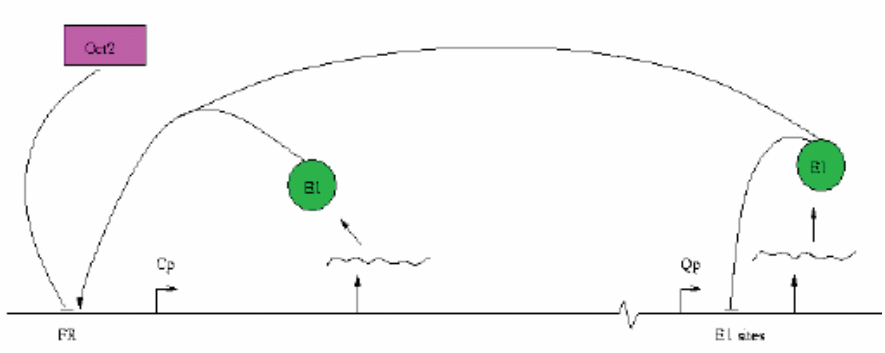

Figure 5.7 Regulation on FR: EBNA-1 & Oct-2 [Courtesy of Maria Werner]

Thus, it is the binding pattern of these two proteins on the FR region that regulate C promoter. While EBNA-1 acts as enhancer of Cp, Oct-2 behaves as inhibitor through blocking neighboring binding sites of EBNA-1. It is believed that Cp can only be ON when at least 8 ENBA-1 have bound to FR [15, 16].

5.4 Presentation of the model

We use statistical mechanics to deal with concentration of molecules inside the cell, a method widely-accepted in chemical simulation, to construct a model able to describe the significant phenomena after primary infection of EBV. Here we look at the general framework of the previous model. Major concern is given to mathematical description of FR region and the C promoter, which provides basis for this thesis work.

Single-side blocking scenario

In principle, the 20 binding sites of EBNA-1 and 20 of Oct-2 molecule in the FR region of the viral genome together gives a total possibility space of 2^{40}. Pure statistics in this problem, i.e. making directionless counting, would be far too time-consuming. In previous work, a reasonable approach of simplification has been made under

assumption that no cooperative binding (molecular cluster is easier to bind than a single molecule) or Double-side blocking (Once Oct-2 is bound to its site, it will stop ENBA-1 from binding at its two neighbouring sites) is involved during the binding of the two molecules.

All cases were classified by n, the number of EBNA-1 and k, number of Oct-2 bound to the region. Oct-2 was assumed to block only the left neighbour site for EBNA-1 to bind and EBNA-1 was accordingly assigned competition for blocking only the right neighbour site of Oct-2. Moreover, to simplify the problem to a deeper level, the competitive binding of the two molecules was performed by combining the neighbouring sites of EBNA-1 and Oct-2, i.e. the total 40 binding sites were treated as 20, where EBNA-1 and Oct-2 are exclusive to each other to bind.

Statistical mechanics applied

We now use statistical mechanics to describe the probability of the system's residence in one binding pattern.

$$w_{nk}([E][O]) = \zeta \bullet [E]^n e^{-n\beta Ee} [O]^k e^{-k\beta Eo}$$

Here n, k are the number of EBNA-1 and Oct-2 molecules bound to FR, repectivesly; [E] and [O] denote the concentration of free E and O; Ee and Eo stand for the binding energy relative to the unbound state. $\zeta = \zeta(n,k)$ gives the number of cases under one particular set of binding pattern. In this simplest scenario, $\zeta(n,k)$ can be obtained through simple combinatorics:

$$\zeta(n,k) = \begin{pmatrix} N \\ n+k \end{pmatrix} \begin{pmatrix} n+k \\ n \end{pmatrix},$$

where N means the total number of sites.

As the ON/OFF activity of the C promoter is believed to depend only on the number of bound ENBA-1: on when $n \geq 8$, i.e.

$$\text{C promoter} \begin{cases} on, & 8 \le n \le 20 \\ off, & 0 \le n \le 7 \end{cases}$$

we need to sum on w_{nk} over k,

$$w_n = \sum_{k=0}^{N-n} w_{nk}$$

$$= \frac{N!}{n!(N-n)!}[E]^n e^{-n\beta Ee}(1+[O]e^{-\beta Eo})^{N-n}$$

So we have total weight as

$$w_{tot} = \sum_{n=0}^{N} \frac{N!}{n!(N-n)!}[E]^n e^{-n\beta Ee}(1+[O]e^{-\beta Eo})^{N-n}$$

$$= (1+[O]e^{-\beta Eo}+[E]e^{-\beta Ee})^N$$

Then w_n divided by the total weight gives the probability of ON over number of bound EBNA-1:

$$P_n = \frac{w_n}{w_{tot}}$$

6 Mathematical requisites: combinatorics

Combinatorics is the area in mathematics dealing with counting under certain constraints. We will build up some paradigm problem together with their solutions so that explanation of the combinatorial problem in our concern becomes clearer.

6.1 Ball-basket: a paradigm problem, Paradigm 0

Before we concern the real challenge in our study, a conceptual paradigm from combinatorics shall be introduced first. It will be used several times when endeavour of solving the problem is made.

Paradigm 0

Suppose we have x balls of the same colour, for example red, and y baskets each signed by a different number, indicating that the baskets are distinguishable while the balls are not. For simplicity $x \geq y$ is assumed here. The task is to distribute these x balls into the y baskets and find all the possible cases with the limitation that any basket shall contain at least one ball and that all balls must be in the baskets.

To solve such problem gracefully is an interesting process with a significant conceptual transformation. Attention shall be paid to the key words 'at least one'. It is based on this fact that we can introduce the approach of inserting a 'separator'. As any basket contains at least one ball, putting balls into y baskets is the same process with inserting $y-1$ separators into the in total $x-1$ blank intervals between each pair of neighbouring balls.

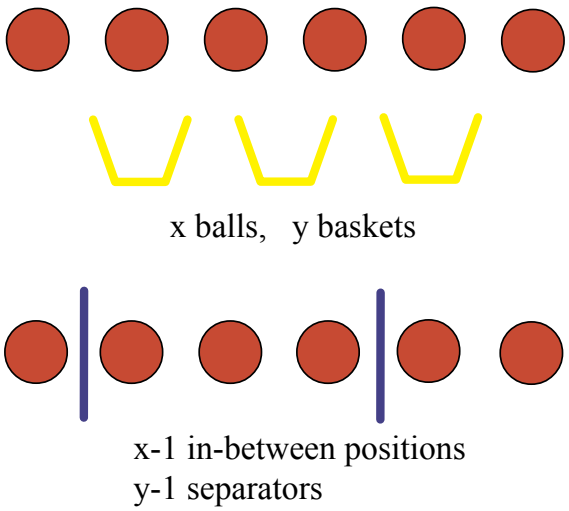

Figure 6.1 Ball-baskets problem using separators

From Figure 6.1, it is not difficult to notice that the number of all possible distributions even needs no algorithm but a closed formula, $\binom{x-1}{y-1}$, denoting that we are choosing $y-1$ out of $x-1$ possible locations for putting separators.

6.2 Generalizing Ball-basket: Paradigm I-III

Paradigm I

Paradigm 0 can be slightly modified to perform tasks of more complexity. One example is to remove the limitation of 'at least one', so that there might be empty basket existing and the problem changes its appearance as distributing x balls into y baskets. However, $x \geq y$ is still assumed.

Although 'at least one' is omitted here, we still try to 'create' such terms by transforming the question, so that the original benefit of using separators become applicable again. We do so by splitting this problem into smaller ones and designing a simple algorithm:

1) Assume a fixed number of empty baskets, q, count all possible cases under this limitation

2) Go through all possible values of q ($0 \le q \le y$).

For step 1), focus is addressed on the rest baskets where distribution of the x balls takes place. First, positions of the q empty baskets must be assured, which includes $\binom{y}{q}$ possibilities. Within the remaining y-q baskets, each must contain "at least one" ball. Utilizing result from Paradigm I, we get $\binom{x-1}{y-q-1}$ kinds of distribution. Therefore total number of possible cases for each value set of (x, y, q) is $\binom{y}{q}\binom{x-1}{y-q-1}$.

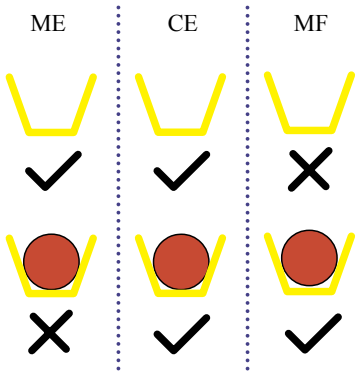

Figure 6.2 Paradigm I, classification of baskets

a) must-be-empty (**ME**) b) can-be-empty (**CE**) c) must-be-filled (**MF**)

ME contains no balls; MF owns at least one ball;

CE can be realized by further classification into ME and MF.

Paradigm II

Furthermore, Paradigm I can be expanded again. What happens if we erase the condition $x \ge y$? The answer lies effortlessly in the solution process of paradigm II. When x is smaller than y, for example $x = y - 1$, there must be at least one empty basket and the case of no empty baskets we considered no longer exists in step 2) of

this new version of question. Thus we can perform the counting by excluding the cases of $x < y - q$, i.e. setting $q_{\min}$, the smallest value of q , to $y - x$, while obviously $q_{\max} = y - 1$.

Paradigm III

This section treats the most general version of Ball-basket problem in our concern. In Paradigm II, we have solved a problem stated below:

- Put x undistinguished balls into y distinguishable baskets. Count all possibilities.

It indeed looks general enough. However, what happens if we know already how many baskets out the total y can be empty? One might optimistically turn to quantity q introduced in Paradigm I, claiming that solution lies in setting $q_{\max}$ to p , the number of can-be-empty baskets and the result is accumulation of $\binom{y}{q}\binom{x-1}{y-q-1}$ over the interval $[\, q_{\min} , p \,]$. This is of course right but only under the following assumptions that every basket can be a candidate of empty baskets. If we already know which p baskets can be empty, we need only choose q out of p, thus the result shall be accumulation of $\binom{p}{q}\binom{x-1}{y-q-1}$ instead.

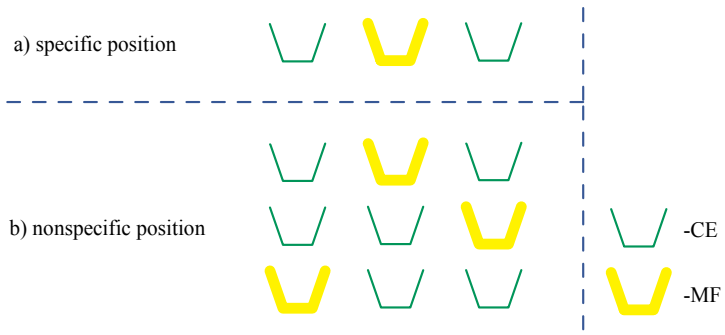

Figure 6.3 Illustration of Paradigm III: only a few baskets can be empty

An algorithm named “ballBasket” is designed according to the discussion above and a logic input quantity is used to determine whether it is the case of $\binom{y}{q}\binom{x-1}{y-q-1}$ or $\binom{p}{q}\binom{x-1}{y-q-1}$.

7 Modeling EBV system

This chapter describes the major algorithm of counting statistical weights of different binding pattern ζ. First we state the necessity of generalizing the previous model and the expectation of result before the work. Then algorithms for different combination of effects are described in detail, including brief presentation of the result of weak cooperativity. Major results are given in the next chapter.

In all cases investigated in the present work, the binding affinity of the proteins bound to FR is set to be -15.45 kcal/mol for EBNA-1 and -12.28 kcal/mol for Oct-2 respectively. The average volume of the host eukaryote cell is approximated as 2×10^{-13} liter. Concentration of proteins presented in the plots is all in M (mol/liter). These are equivalent with parameters used in [17].

7.1 More than single blocking

Assumptions in previous model were made for limiting computational costs. However, the more phenomena a model includes, the better prediction it can potentially make. We want to build a more exact model. Though requiring more complex numerical calculations of ζ than the simple combinatorics in section 5.4, this is rewarding for both theoretical research and designing experiments.

It is possible that one bound molecule/complex to the DNA blocks both its neighbouring sites, on each side, since the separation between the binding sites are small. Thus, we shall give a model able to simulate O's blocking on both neighbouring E sites. Moreover, EBNA-1 is known to be able to bind cooperatively to other binding sites on the EBV genome [18,19,20]. Though the cooperative effect at the FR region has not been tested directly, we assume this could happen. All cases that need to be simulated are concluded in Table 7.1.

Cases	Cooperativity	Single blocking	Double blocking

1	No	Yes	No
2	No	No	Yes
3	Yes	Yes	No
4	Yes	No	Yes

Table 7.1 Combinations of effects for investigation

For studying cooperativity and its relationship with competition, we want to vary the cooperative energy see the consequences. In particular, we want to compare our result, an example of multi-site system with previous study on simpler systems with just several binding sites, where blocking is less likely to happen.

7.2 Competitive binding of EBNA and Oct

It has been pointed out in previous sections how important competitive binding might be involved in the viral survival strategy. Here we begin to describe this molecular mechanism quantitatively using results of the established combinatorial paradigm.

7.2.1 Double blocking

Oct-2 is now assumed to be able to block not only left but also its right neighbouring site of ENBA-1. As the left neighbour competition is already included, shrinking total sites from 40 to 20, we need merely focus on the competitive binding of Oct-2 on the right neighbouring ENBA-1. Based on the mechanism that will be discussed later for describing blocking, a natural expansion of this method can be applied to cooperative binding.

Same for any O excluding E as right neighbour, any E must not have an O on the left. In other words, there shall be “at least one” empty binding site Φ between any “OE” segments when efforts to specify a whole sequence of E, O, Φ are made. This binding feature can easily recall us of Paradigm III in the balls-baskets problem. There is an obvious analogy of the intermediate position of every two molecules in S1 being

baskets, while the empty binding sites play the role of balls. However a slight extra considerance shall be pointed out for this analogy that there must be two more "baskets". In a 20 sites sequence, the first several sites or the last ones could be either taken up by molecules or empty, giving birth to two more can-be-filled baskets.

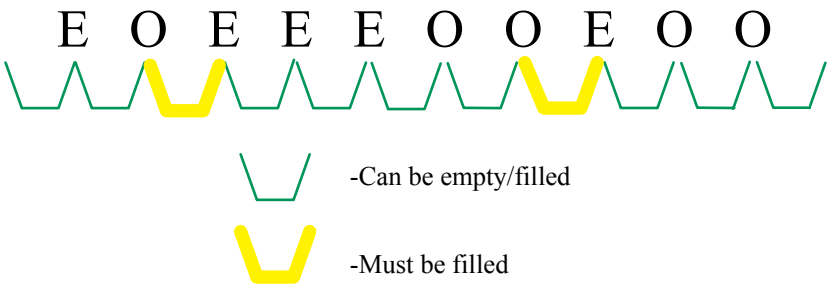

Figure 7.1 Front, end, in-between positions as baskets, Φ as balls

S1 Element	Basket Type
front	can be filled/empty
end	can be filled/empty
OE	must be filled
EO	can be filled/empty
EE	can be filled/empty
OO	can be filled/empty

Table 7.2 Basket rule for "single blocking"

Thus, we can use such procedure to "produce" a complete FR region with molecules bound:

1) Construct a "pure molecule sequence" (S1) made up only by E and O.

2) Find all "OE" segments in this sequence

3) Treat the positions between O and E in such segments as must-be-filled baskets; put Φ as balls into them. Other spaces between molecules in any "EE", "EO" and "OO" segments are treated as can-be-empty baskets.

Denote the total number of must-be-filled baskets by z, we have in total x balls, n+k+1 baskets. Then we can set the largest possible number of empty baskets to n+k+1-z and replace the term of $\binom{y}{q}$ by $\binom{n+k+1-z}{q}$ saying there are only n+k+1-z positions to fill for empty baskets.

The remaining work then is to count the number of all possible cases of pure E-O sequence (S1) for each fixed value of z. Partly due to the "inclusive nature" of our expectation of the model, and partly because of practical need, we choose to distinguish between the OE and EO segments, denoted by z_{OE} and z_{EO} respectively. A helpful thing is that soon we will see such classification yields a surprisingly regular counting method.

Two important natural characteristics can be drawn out of case studies of specifying sequences with z_{OE} and z_{EO}. Firstly, z_{OE} and z_{EO} can not be independently arranged. As seen the fig below, for each fixed z_{OE} there exist only three possible values z_{EO} at most, z_{OE} -1, z_{OE} and z_{OE} +1. And this provides an advantage of shrinking the searching space that we always strive to find. Secondly, for each (z_{OE} , z_{EO}) value pair, emergence of at most 4 balls-baskets problem is found, whose solution has already been established in former statements.

7.2.2 Ball-basket revisited

A fixed number of z_{OE} and z_{EO} means also specified backbones of the pure n+k long E-O sequence. Taking typically the case of z_{OE}=1 for illustration, we can at first put solely one OE segment in centre. If z_{EO}=0, only once backbone exists: <u>OE</u>, the rest O must all be on the left and the remaining E on the right. Otherwise any E on the left or O on the right would introduce an unwanted EO segment. For z_{EO}=1, the backbone might be E<u>OE</u> or <u>OE</u>O while z_{EO}=2 holds only E<u>OE</u>O.

	Number of neighborhood	
Backbone	OE	EO
$b_E\ b_O$	0	1
$b_O\ b_E$	1	0
$b_E\ b_O\ b_E$	1	1
$b_O\ b_E\ b_O$	1	1
$b_E\ b_O\ b_E\ b_O$	1	2
…	…	…

Table 7.3 Illustration of Backbones

With foundation of a framework, referred as backbone above, the last step of establishing the pure E-O sequence is to insert the remainder of E and O into allowed positions. As stated before, this is again the balls-basket problem. Take the symbol E in the backbone as baskets (b_E) for E to fill and backbone O (b_O) for O. This time, it is Paradigm 0 that shall be applied, as all b_E, b_O must be filled with at least one corresponding molecule.

As this method will be used and modified for later sections, we name this method ‘backbone description’ and summarize it as blows:

1) Go over all possible backbones (S0) and complete 2) to 6) for each backbone.

2) Use paradigm 0: E, O molecules as balls; b_E as baskets for E, b_O for O.

3) 'Pure molecule sequence' (S1) made up only by E and O generated.

4) Find and record the number of all 'OE' segments in S1.

5) Use paradigm III: Φ as balls; intermediate positions in ‘OE’ segments as must-be-filled baskets; other spaces in 'EE', 'EO', 'OO', front and end as can-be-empty baskets.

6) Final complete sequence of FR (S2) obtained.

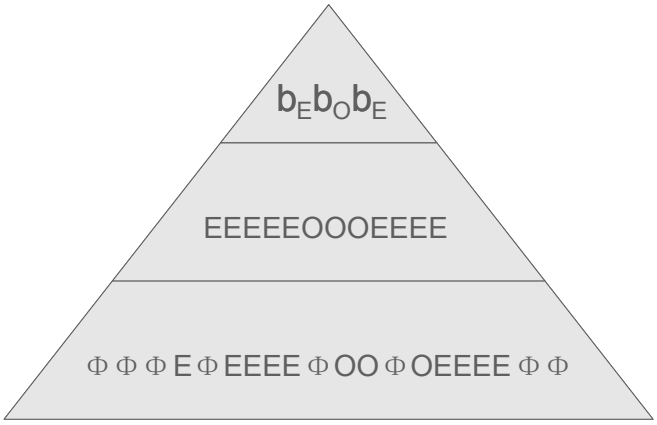

Figure 7.2 Backbone description: from backbone to FR.

7.3 Cooperative binding

Cooperativity includes many cases to consider. To simplify the treatment so that algorithm of executive efficiency is possible to make, we consider only two extreme case of cooperativity: weakest and strongest cooperative binding. Here cooperativity is tested for both EBNA-1 and Oct-2.

7.3.1 Weak cooperativity

The weak cooperativity is defined as follows:

1. Cooperative binding happens only between neighboring molecules of the same type.
2. Once an EBNA or Oct cooperatively binds to one side of neighboring counterpart, it won't be able to cooperatively bind to the other side.

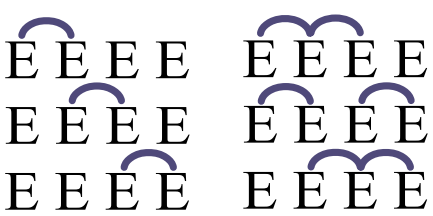

Figure 7.3 Schematic of weak cooperativity

Though this description might not be realistic due to introduction of Requirement 2, it is the easiest to implement computationally and it indicates some useful information about the cooperative effects. The computational idea is simply treating cooperative binding EE and OO as new elements when we try to construct the whole binding patterns of FR region. Thus, we need to introduce another two quantities n_1, k_1, each representing number of cooperative bound elements of EBNA and Oct. For simplicity, we refer cooperative bound elements to C for EBNA and D for Oct.

The remaining task is then to put all elements into the empty FR region. We need to

1. Sort all possibilities of non-cooperative binding elements with the cooperative ones, which gives, $\binom{n+2n_1}{n}\binom{k+2k_1}{k}$.

2. Sort EBNA with Oct, $\binom{n+n_1+k+k_1}{n+n_1}$

3. fix positions of all taken binding sites, choosing $n+n_1+k+k_1$ out of $N-n_1-k_1$, $\binom{N-n-n_1-k-k_1}{n+n_1+k+k_1}$

To note that this method uses no 'backbone description' of the sequence structure, which will be inevitable to be applied when cooperativity and double blocking are both involved. Here cooperative energy of the molecules is expressed relative to binding energy of respective molecule.

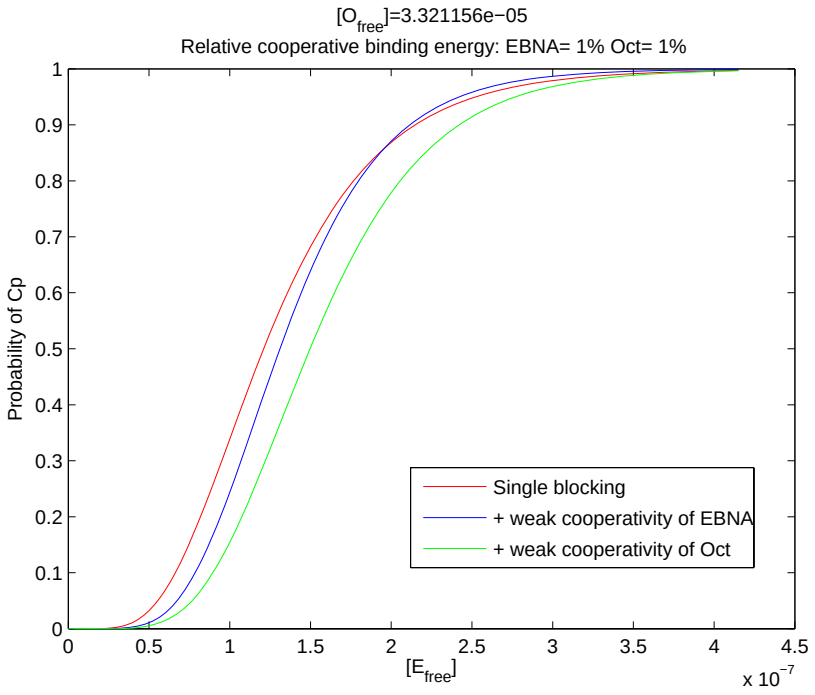

Figure 7.4 Weak cooperativity + single blocking

From the picture above, we can observe what cooperative binding, even just in its weakest manner, will bring. When EBNAs can help each other to bind to FR, the curve becomes sharper than in single blocking scenario, making the whole system behave more like a switch. Conversely, cooperative binding affinity of Oct will add to influence of blocking; rising of probability from zero will be postponed.

Except the information provided by the weak cooperative scenario, there is still a critical problem. As we have treated the cooperative binding EBNA-1s and Oct-2s as different elements with the non-cooperative binding molecules, the algorithm will fail due to repetitive counting if we want to bring this model back to the single blocking scenario by simply setting the relative cooperative binding energy to zero. Unfortunately one of our major interests is to understand quantitatively how the 'reactivity velocity ' of this complex genetic switch is affected by cooperative binding of molecules. This demands consistency of the model at transition from non-cooperative scenario to cooperative ones. Our model needs reconstruction.

7.3.2 Strong cooperativity

By strong cooperativity, we mean the following consideration:

1. Cooperative binding happens only between neighboring molecules of the same type.
2. Any single molecule is able to cooperatively bind to DNA with both its left and right neighbors.
3. Any molecules of the same type bound to the DNA neighboring each other are definitely cooperatively bound as a whole.

Here the term of 'cooperative binding molecules' equals 'neighboring molecules'.

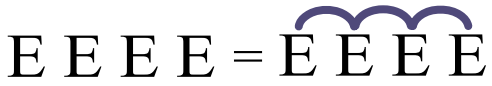

Figure 7.5 Schematics of strong cooperativity

Unlike the weak cooperativity scenario, we can't treat cooperative binding molecules as different elements constructing the FR region for reasons given in previous section.

It is somewhat cumbersome to develop a method that provides the number of cases specified by the n, k, n_1, k_1, as we must know not only n and k, but also the neighboring relationship inside such settings of n and k, which is highly dependent on the structure of sequences. Essence of this problem lies in the nature of the core method applied--combinatorics. To count is a process of gathering information about something, the counting object, under certain conditions. These conditions, after the counting, become representatives characterizing the object while other characteristics of the object are hidden in the number and no longer visible. So it is fair to conclude more precisely that combinatorics is a process that gathers and meanwhile loses information of the counting object. The more information we need about the object, the closer we must be to the object. Some may then suggest such straightforward alternative:

Go over all possible sequences and record

1) n, number of ENBA-1 bound to the FR
2) k, number of Oct-2 bound to the FR

3) n_1, number of neighborhood of all EBNA-1 on FR
4) k_1, number of neighborhood of all Oct-2 on FR

and make statistics of all sequences counting the number of cases under each settings of n,k,n_1,k_1.

However, this violates our initial purpose to avoid such time-consuming statistics of a size of order 3^{20}. Therefore we must find some way that combines advantages of both the detailed sequence description and the efficiency provided by combinatorics. This method of description should be not only detailed enough to so that necessary information of the sequence can be acquired but also sufficient enough so that going over all possible sequences can be faster.

Fortunately such approach exists for this simple system concerning only two types of molecules. We can choose to count the number of neighboring molecules of the same type rather than to assign values of n_1 and k_1 on the first place, with help from the backbone description.

As backbone description involves two steps of constructing the sequence made up of E, O and Φ, the first step of creating a pure E-O sequence (S1) provides some information of the neighboring structure. Though it is only neighboring structure of the pure E-O sequence, we can modify the process of inserting Φ to provide information we need.

When generating S1 from a backbone, we know precisely how many E, O and how many backbone-baskets are there for Es and Os, which are potential candidates for neighboring relationship. And the only possibility to split these potential neighbors up is to insert a number of empty binding sites Φ when producing S2.

For example, in a backbone of "$b_O b_E b_O b_E b_O$", where number of backbone-baskets for E and O are respectively 2 and 3, n Es must be put into 2 baskets. In each of the 2 E-backbone-basket, all molecules contained must be neighboring each other, giving number of such neighborhood as n-3 for E and k-4 for O.

Number of backbone baskets		Number of neighborhoods	
b_E	b_O	EE	OO
1	1	n-2	k-2
2	1	n-3	k-2
2	3	n-3	k-4
a	b	n-1-a	k-1-b

Table 7.4 Number of potential neighborhood in backbones

We shall record these two numbers and then decide how to reduce them by inserting empty binding sites. Inspiration then comes again from the procedure of ball-basket algorithm. When solving Paradigm I, we created room for algorithm of Paradigm 0. That is, split the problem by q, and use results from Paradigm 0. This idea of introducing a fixed number of can-be-empty baskets can lead us to more strict description—a fixed number of must-be-empty baskets, m. This can be easily done by ball-basket algorithm, setting the total number of baskets to y-m. By doing so, we are able to divide the n_1 E neighborhoods and k_1 for O into two non-superpositional categories of baskets for Φ: 1) must-be-filled and 2) must-be-empty. The algorithm then becomes:

1) Choose v out of n_1, w out of k_1, setting them as well as must-be-filled baskets. At the same time, remaining n_1-v and k_1-w are considered as must-be-empty baskets.
2) Insert Φ as balls according to the "basket rule".
3) Go over all possible values of v (0~n_1), w (0~k_1).

This basket rule is summarized in Table 7.5 below.

Sequence Element	Basket Type	Number
OE	can be filled/empty	z_{OE}
EO	can be filled/empty	z_{EO}

EE	must be filled	v
EE	must be empty	n_1-v
OO	must be filled	w
OO	must be empty	k_1-w

Table 7.5 Basket rule for “strong cooperativity”

7.4 Cooperative binding + double blocking

Discussions until now enable us to test the combined effects of both Double-side blocking and cooperativity. Weak and strong cooperativity are both considered.

7.4.1 Weak cooperativity + double blocking

Weak cooperativity is not as straight-forward to implement as strong cooperativity when blocking must be taken into account. The reason is apparent: it must be fit into the framework of 'backbone description' so that current algorithm from the Double-side blocking can be partly useful. Recalling the 3 steps of treating weak cooperativity, we are able to notice that some of the tasks of these steps are already finished in the 'backbone description', namely Step 2 and 3. This finding necessitates the introduction of a similar step 1 into the 'backbone description', which is not difficult to implement.

The idea is almost the same as the way we consider double blocking. The core and only difference is that we sort the non-cooperative binding elements with the cooperative ones, count number of all the possibilities and then consider them as unified elements so that developed double blocking algorithm can be used without any more modifications.

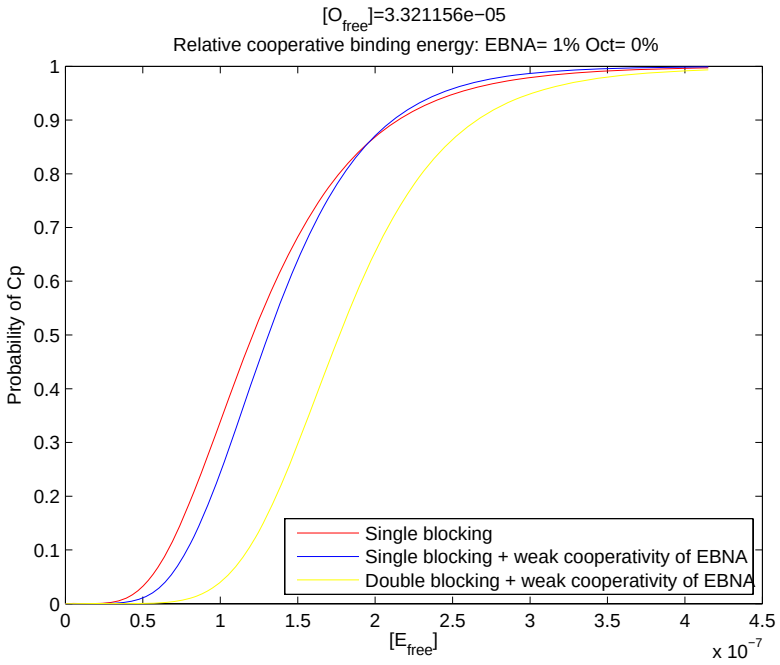

Figure 7.6 Weak cooperativity + double blocking

Comparing the blue and yellow curves in Figure 7.6, it is obvious that double blocking has drawn the switch to happen at higher concentration of EBNA-1 just like the case when cooperativity of Oct-2 is considered.

7.4.2 Strong cooperativity + double blocking

The way to carry out combination of strong cooperativity and double blocking is effortless. The only thing we need to do is to change the "basket type" of OE segments to must-be-filled, compared with treating strong cooperativity only. A detailed description of basket rule for this case is again given in a table.

Sequence Element	Basket Type	Number of baskets
OE	must be filled	z_{OE}
EO	can be filled/empty	z_{EO}
EE	must be filled	v
EE	must be empty	n_1-v

OO	must be filled	w
OO	must be empty	k_1-w

Table 7.6 Basket rule for “strong cooperativity + double blocking”

8 Results

As the model of weak cooperativity is not consistent at the transition point from the non-cooperative case to cooperative one, and as cooperativity of EBNA-1 but none of Oct-2 is already suggested, we discuss here only results of strong cooperativity of EBNA-1.

8.1 Sensitivity of the switch

Figure 8.1 gives a 3D plot of the probability of C promoter for 4 different combinations of strong cooperativity and blocking effect over concentration of EBNA-1 and Oct-2. It is not difficult to notice that the consequence of adding strong cooperativity of EBNA-1 is generally consistent with that of adding weak cooperativity. The only obvious difference is that the probability of the strong cooperativity scenario rises from zero earlier than that in the case of weak cooperativity (in Figure 7.5 the curve with weak cooperativity crosses the one of single blocking rather than being drawn to lower concentration of EBNA-1), which verifies our definitive difference between weak and strong cooperativity.

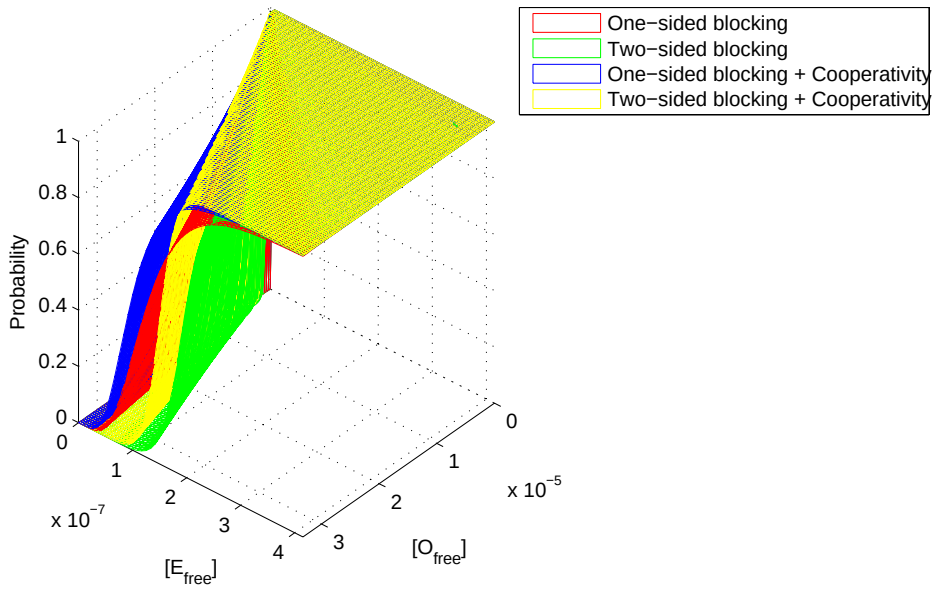

Figure 8.1 Probability of C promoter, 3D plot

The switch of Cp becomes more sensitive when strong cooperativity or blocking effect or both of them are taken into consideration. However, when looking into more details of the strong cooperative effects in next chapter, we will see this consistent general trend can be differentiated into complex reasons.

8.2 Cooperativity V.S. competition

Results of Section 8.1 interest us to consider more about the system. This section deals with the interrelation between the competitive and cooperative effects. We develop another model concerning pure cooperativity to differentiate phenomena caused by the two effects and use Hill coefficient as measure of the degree of cooperativity of EBNA-1. Again we limit discussion of cooperativity to strong scenario only.

Pure cooperativity

Pure cooperativity is first discussed here to define Hill coefficient in our case and to confirm whether our result is consistent with those of previous studies where only several binding sites are involved in the genetic switch.

Hill coefficient is defined as the derivative of the logarithm of the ratio of occupied promoter over the free operator versus logarithm of the concentration of the free transcriptional factor. Suppose we have the following reaction.

$$nA + O \rightleftharpoons A_nO$$

where A stands for the transcriptional factor, O for the operator on the DNA. In equilibrium, the reaction constant K is defined as

$$K = \frac{[A_{free}]^n[O]}{[A_nO]}$$

Considering $[O_{total}] = [O] + [A_nO]$, the fractional occupancy, also here known as the probability of the operator to be ON is

$$P = \frac{[A_n O]}{[O_{total}]}$$

We then have

$$\lg(\frac{[A_n O]}{[O]}) = \lg(\frac{P}{1-P}) = n \cdot \lg([A_{free}]) - \lg(K)$$

And Hill coefficient is expressed as

$$\frac{\partial \lg(\frac{P}{1-P})}{\partial \lg([A_{free}])} \quad (P = 0.5)$$

Apply above results to our EBNA-FR system, the following result is obtained.

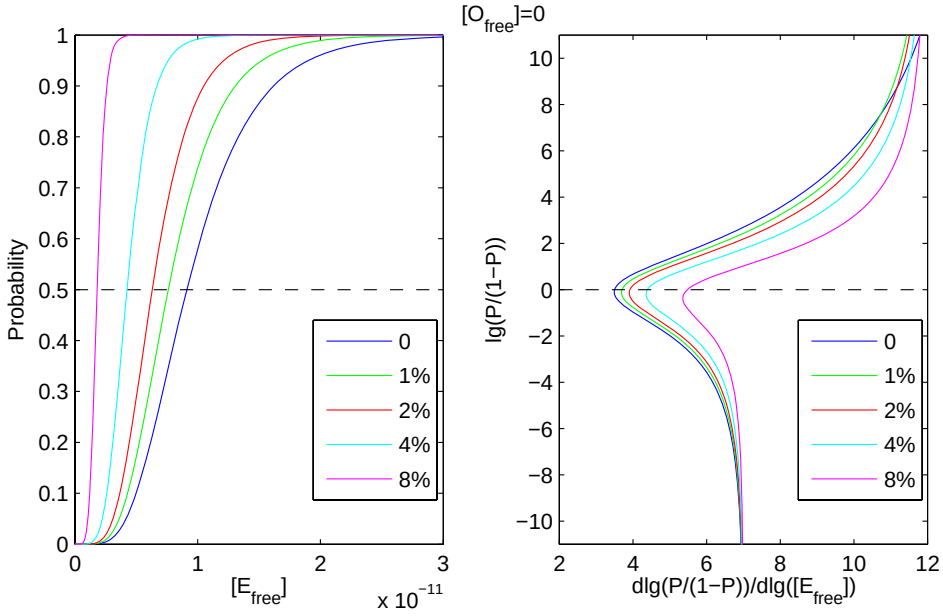

Figure 8.2 Pure strong cooperativity of EBNA-1

From Figure 8.2, we can see that Hill coefficient is usually defined at half saturation (P=0.5, dashed line in plot). It is close to this point that the minimal value of the quantity $\frac{\partial \lg(\frac{P}{1-P})}{\partial \lg([E_{free}])}$ is obtained. With growth of the relative cooperative energy of EBNA-1 (relative to the binding energy of a single EBNA-1 molecule), an increase of

the Hill coefficient is clearly shown. This result is in consistence with previous study on viral systems consisting of only several binding sites.

Strong cooperativity + single blocking

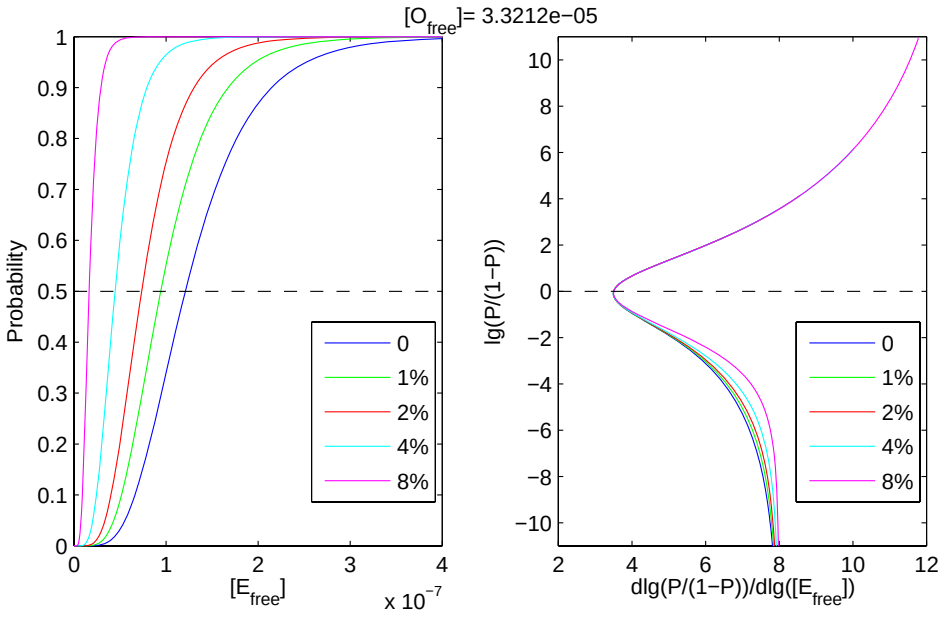

Figure 8.3 Strong cooperativity of EBNA-1 + single blocking

When single blocking is involved, the increase of Hill coefficient with the growth of cooperative energy of EBNA-1 vanishes. This can be explained by a relatively simple notion of "effective binding sites". After introduction of blocking, effectively speaking, the actual binding sites to which EBNA-1 can bind are reduced. The result of this decrease in effective binding sites of EBNA-1 can be revealed by analogy between the number of effective binding sites of EBNA-1 in the single blocking scenario and the total number of binding sites in the case of pure cooperativity.

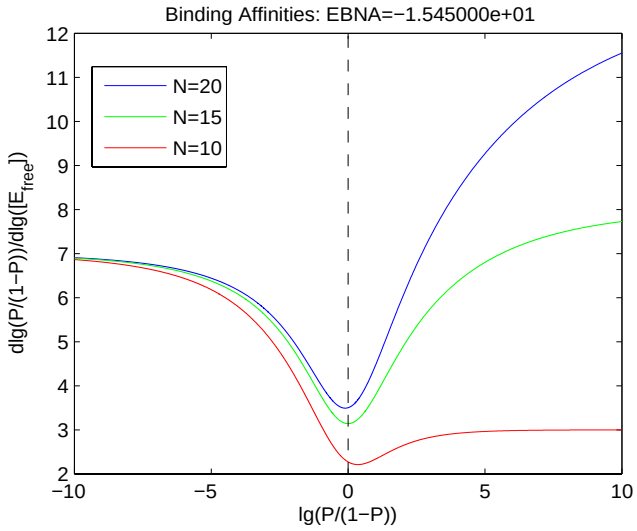

Figure 8.4 Effective binding sites of EBNA-1

Shown in Figure 8.3, in pure cooperativity scenario, decreasing total number of binding sites gives smaller Hill coefficient. High [O] means many of the 20 binding sites are taken by Oct-2, inhibiting neighboring sites and split the whole sequence into pieces to bind EBNA-1. The course of cooperative binding can be understood as combination of two steps. EBNA-1 molecules try to bind first, and those bound as neighbors find ways to cooperate, forming a state of lower free energy. When number of neighborhoods for EBNA-1 is reduced much by existence of Oct-2, the degree of cooperativity will not be as large as in the pure cooperative scenario. In high [O], there can be little neighboring EBNA-1 bound to FR, which makes the raise of cooperative energy in vane. Moreover, the fact that strong cooperative binding at neighboring sites is defined equivalent to the neighborhood itself drives such influence from blocking to the maximum.

Strong cooperativity + Double blocking

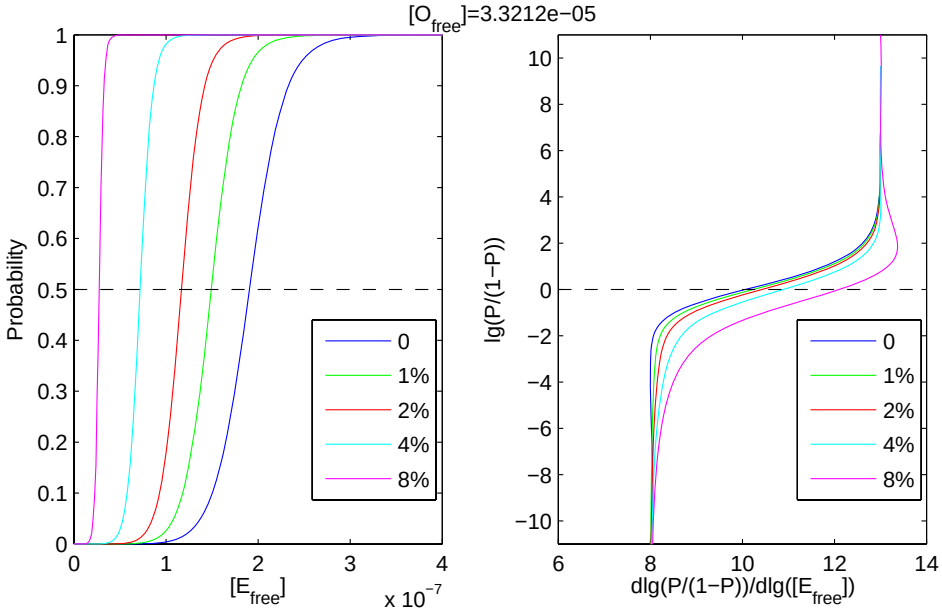

Figure 8.5 Strong cooperativity of ENBA-1 + double blocking

Double blocking introduces more complex result than what single blocking has brought. Minimum of the quantity $\frac{\partial \lg(\frac{P}{1-P})}{\partial \lg([E_{free}])}$ is no longer obtained at half saturation. But the rise of switch sensitivity with growth of cooperative energy of ENBA-1 is still observed.

9 Discussion

It can be concluded from the results that in this specific example of multi-binding-site genetic switch, competitive effects are more important than cooperative effects. Qualitative properties of the Hill curve are dominantly determined by the type of competitive bindings but will not change much under a relatively large alteration of cooperative binding energy.

To fit ourselves in a more general context, say the field of systems biology as a whole, it is not difficult to observe from methodology of this work that this field of research is accordingly initiated by biology rather a systematic description manner. It is already a time in which biologists call mathematicians and physicists for explaining the enormous amount of experimental data. When we are pleased at the serviceability of modeling in living systems, we must also notice that what we call system virology seems only able to model viral behaviors that are critical for infected cells and already recognized in the biology community. That is the task to identify whether a protein or mechanism is essential for a specific problem will almost always turn to the biologists.

A feasible way of making modeling tools more powerful at predicting behavior of living systems may be one of the engineering methodologies—to integrate. By making libraries of useful modeling so that previous successful physical or pure mathematical models can be remembered, shared, improved and integrated for problems at larger scale. Thus, statistical physicists that are ambitious to make more contribution in systems biology shall not only be skillful at what is required for a traditional physicist but also open to absorb artifice of chemical engineering and software engineering.

References

[1] N. Rosenfeld, J. W. Young, Uri Alon, P. S. Swain, M. B. Elowitz, Science 307, 1962 (2005)

[2] M. Stetter, G. Deco, and M. Dejori, Artificial Intelligence Review 20:75 (2003).

[3] Stefan Bornholdt, Science 310: 449-451 (2005)

[4] Lustig, A. J. Levine, J. Virol. 66: 8: 4629

[5] L. Bintu, N. E. Buchler, H. G. Garcia, U. Gerland, T. Hwa, J. Kondev, T. Kuhlman, and R. Phillips, Curr Opin Genet Dev 15, 116 (2005).

[6] L. Bintu, N. E. Buchler, H. G. Garcia, U. Gerland, T. Hwa, J. Kondev, T. Kuhlman, and R. Phillips, Curr Opin Genet Dev 15, 125 (2005).

[7] D. A. Wysokenski and J. L. Yates, Journal of Virology 63, 2657 (1989).

[8] H. Zetterberg, C. Borestrom, T. Nilsson, and L. Rymo, International Journal of Oncolocy 25, 693 (2004).

[9] T. Nilsson, H. Zetterberg, Y. C. Wang, and L. Rymo, Journal of Virology 75, 5796 (2001).

[10] J. Cruickshank, K. Shire, A. R. Davidson, A. M. Edwards and Lori Frappier, J. Bio. Chem. 275, 29, 22273 (2000)

[11] R. F. Ambinder, W. A. Shan, D. E. Rawlins, G. S. Hayward, and S. D. Hayward, Journal of Virology 64, 5, 2369 (1990)

[12] C. P. Verrijzer, J. A.W.M.van Oosterhout, W. W.van Weperen, and P. C.van der Vliet, EMBO 10:10:3007 (1991)

[13] C. P. Verrijzer, M. J. Alkema, W. W.van Weperen, H. C.van Leeuwen, M. J.J.Strating, and P. C.van der Vliet, EMBO 11:13:4993 (1992)

[14] P. C. Shah, E. Bertolino, and H. Singh, EMBO 16:23:7105 (1997)

[15] J. Almqvist, J. Zou, Y. Linderson, C. Borestrom, E. Altiok, H. Zetterberg, L. Rymo, S. Petterson, and I. Ernberg, Journal of General Virology 86, 1261 (2005).

[16] S. Malin, Y. Linderson, J. Almqvist, I. Ernberg, T. Tallone, and S. Petterson, Nucleic Acids Research 33, 4618 (2005).

[17] M. Werner, I. Ernberg, J. Zou, J. Almqvist, and E. Aurell, submitted to BMC Systems Biology.

[18] H. Summers, J. A. Barwell, R. A. Pfuetzner, A. M. Edwards, and L. Frappier, Journal of Virology 70, 1228 (1996).

[19] Bochkarev, J. A. Barwell, R. A. Pfuetzner, E. Vochkareva, L. Frappier, and A. M. Edwards, Cell 84, 791 (1996)

[20] J. Cruickshank, K. Shire, A. R. Davidson, A. M. Edwards, and L. Frappier, J. Bio. Chem. 275:29, 22273 (2000)

[21] Hannon GJ, Nature 418:244:51 (2002)

[22] Epstein. M. A. Barr, Y. M. & Achong. B.G., Lancel 15, 702 (1964)

[23] L. S. Young & A. B. Rickinson. Nature Rev Cancer 4, 757 (2004)

[24] E. R. Leight & B. Sugden, Rev. Med. Virol. 10:83 (2000)

[25] M. Ptashne, A genetic switch (3rd edition), Cold Spring Harbour Laboratory Press, 2005

[26] M. Ptashne and A. Gann, Genes and Signals, Cold Spring Harbour Laboratory Press, 2002

[27] Peter J. Cameron, Combinatorics: topics techniques, algorithms, Cambridge Press, 1994

[28] Kim Sneppen, Giovanni Zocchi, Physics in molecular biology, Cambridge Press, 2005

[29] Richard Dawkins, The Selfish Gene, Oxford University Press, 1989

[30] http://www.accessexcellence.org/RC/VL/GG/central.html

Acknowledgements

First of all, I would like to express my sincere thanks to Mikael Fogelström and Prof. Göran Wendin for accepting me into the Mater Program of Nanoscale Science and Technology, providing a rich and stimulating study experience, and giving me the opportunity to complete my master thesis at a different university, the extra administrative procedure brought by such transfer notwithstanding.

I warmly thank my supervisors Maria Werner and Prof. Erik Aurell for their patience and exceptional scientific guidance, which has lead to this thesis. A special thank goes to Aymeric Fouquier d'Herouel and his pet family of fruit flies, who, together with Maria and Erik, create the excellently enjoyable working environment for me.

I greatly acknowledge all colleagues in the KTH Computational Biology building. I did benefit considerably from the talks in the seminars of SBC, CBN, CBB and from all other casual conversation with all nice people here.

My great heartfelt gratitude goes to my beloved parents, Zhengpin Ren and Zimin Zhu, for their being a most patient audience and advisor to me when difficult problems arise and for their great confidence in me all through my years abroad. I would like to express my loving appreciation to my girlfriend Shuo Zhang, who, as always, supports my work by all means possible.

Zhu Lizhe

Stockholm, 2007