STATISTICAL SIGNAL PROCESSING APPROACH TO *DNA* REPAIR

***Ram Sever and Hagit Messer*[1]**

School of Electrical Engineering, Tel Aviv University, ISRAEL

ABSTRACT

Signal processing, and especially statistical signal processing, is a field in which generic tools for modeling, analysis and processing of signals are developed. Traditionally, it has been used in technology, and most modern technological systems apply advanced signal processing. However, the post-genomic era introduces challenges which, from a signal processing point of view, may lead to new understanding and promising results. We suggest to apply statistical signal processing tools to the problem of DNA repair, where nature operates as a master engineer. The DNA repair process consists of small machines (proteins, enzymes), which continuously transmit and receive signals from each other. The system regulates its operation; it has feedback loops and backup paths. We suggest modeling the components of the DNA repair system by a probability Markov state diagram.

1. INTRODUCTION

Recent years have introduced new technologies, such as DNA micro array chips, which enable to gather an enormous amount of information in the area of cell biology. One of today's science important challenges is to analyze and understand it. One way to achieve this is through biological modeling. Modeling is not new in the field of molecular biology. Regulation networks and global system approaches have been around for many years. However, modeling became more popular once the sequencing of the human genome was completed, and effort was turned to understand what is written in it[1][2].

The starting point to our work is the following question: How can the success or failures of the DNA repair mechanisms be quantified? How does one measure DNA damage? To answer the above questions, we have developed a mathematical/statistical model to describe the enzymes' operation[3]. Our model can be used to translate bio-chemical information about the enzymatic reaction, such as: the concentrations of all components, the diffusion coefficients, the rate laws and the rate constants, to probabilistic information. We propose to use this model to quantify, in probabilities terms, the various repair or damage events. Our approach can lead to the ability to simulate the complete DNA repair process of an *E-coli,* and its prediction can be verified by experiments. However, it requires knowing the biochemical parameters of all enzymes involved in the DNA repair process, which are not yet available. Since no high-throughput procedure for getting these parameters exists, this is an ambitious, long term task.

2. GENETIC COMMUNICATION

The main task in any communication system is to transmit and receive information between different points on a given physical media reliably. Communication systems are often modeled by a simple block diagram of three components which includes a transmitter, a receiver and a channel.

Communication systems have been intensively studied by engineers, during the passing decades, in order to determine what the best way to transmit information over them is. In 1948 Shannon has laid in his paper: “Mathematical theory of communication”[4] the very first foundations of Information Theory and modern digital communication. Shannon determined the term “channel capacity”, which is the maximum mutual entropy between the input and output sequences, as a measure for the maximum amount of information which can be reliably passed over a given channel.

Living cells face a task similar to communication systems. In order to function, they must continuously deal with both external and internal hazards which threaten to damage the integrity of the cell’s DNA. A “success” in terms of molecular biology may be considered as the cell's ability to reproduce, and to pass its genetic information to its offspring. Figure 1 presents a simplified model for the cell’s DNA life cycle in terms of information flow.

[1] School of Electrical Engineering, Tel Aviv University, Tel Aviv 69978, ISRAEL. E-mail: rams, messer@eng.tau.ac.il

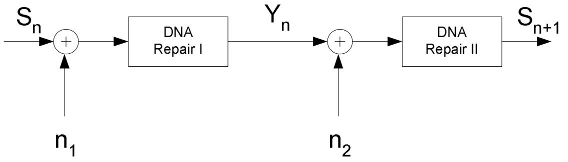

Figure 1: Cells life cycle. S_n is the DNA of the cell at generation n; Y_n is the DNA pre the replication process; S_{n+1} is the DNA at generation: n+1.

The input signal S_n represents the cell's genome at generation-n, and the output signal S_{n+1} represents the cell genome at generation n+1, after the replication process. The intermediate signal Y_n represents the DNA before the replication. In general, $S_n \neq Y_n$ because during the cell's life DNA damage, denoted by n_1 is introduced, and is partially corrected by the first DNA repair block. Replication errors are represented by the additive noise process n_2 and are partially corrected by the second DNA repair block.

To keep complexity limited, we concentrate on *E. coli*, in which the damage repair system consists of five main DNA repair mechanisms[5]:

1. Nucleotide excision repair (NER).
2. Base excision repair (BER).
3. Mismatch repair (MMR).
4. DNA repair by damage reversal (DR).
5. Recombination repair.

Our model suggests putting the various DNA repair mechanisms into two groups. The first is a general repair process, which is composed of repair mechanizms as the BER and the NER. These mechanizms operate along the cell life cycle and are responsible to correct any damage to the DNA as a consequence of both internal and external lesions. The second is the mismatch repair mechanism, which is responsible to correct DNA polymerase mismatches. The two blocks describe different repair processes, with different strategies and target goals. The various DNA damages are introduced by two additive noise signals which may have different statistical models.

3. THE SUGGESTED MODELING

To be able to use the block diagram of Fig. 1, one needs to specify the different component of it. The "noise" processes n_1 and n_2 are to be modeled by some stochastic processes which characterize the random nature of the DNA damages. In this paper we concentrate on modeling the two DNA repair blocks. Bearing in mind that a DNA repair operation is of a random nature (it can succeed or not), our aim is to translate a bio-molecular DNA repair process into a simplified binary channel between the binary input/output: damaged/non-damaged DNA, by assigning it the appropriate transition probabilities. In that setting, the transition probability from "damaged DNA" to "non-damaged DNA" is the probability of a successful repair, while the transition probability from, say, "non damaged DNA" to "damaged DNA" is the probability of miss-repair. These probabilities depend on the operation of the enzymes in the different DNA repair mechanisms and on the inter-relations between them.

We suggest a new approach[2] to statistically model enzyme operation. The use of this approach for the DNA repair mechanisms enables one to calculate the system's performance in terms of probabilities. Once our modeling is complete, it can be used to quantify, at any given time interval, the different transition probabilities in the "damaged DNA"/"non-damaged DNA" binary channel.

The transition probabilities are expressed as a function of the physical parameters, as the enzymes' concentrations and their kinetic velocity constants. This provides an important analysis/synthesis tool. For example, it can be used to study the effect of the level of certain enzymes on the resulting DNA repair.

3.1 Markov model for enzymes operation

Consider a simple enzyme system which catalyzes a single chemical reaction that transforms the substances-S to the products-P. The above enzyme operation can be described by the following equation:

$$E + S \underset{K_{-1}}{\overset{K_1}{\rightleftarrows}} ES \xrightarrow{K_2} E + P \qquad (1)$$

Where K_1, K_{-1} and K_2 are the rate velocities of the process. Note that both the enzyme and the substance can be in number of states; the enzyme can be found in one of two different states, while the substance can be found in one of three states. Both the enzyme and the substance move from one state to the other, as described by (1). For example: the enzyme is constantly moving between its two states as it catalyzes the production of more and more products. However, some states are final, such as the "product state", since once the substance has been turned into a product, it remains as such. We can use Markov state diagrams to describe the system states and the transitions between the states. We have constructed two Markov diagrams, one for each component of the enzymatic system (see Fig. 2).

[2] An approach similar to ours has been taken in the work of Hassibi et. al[6] which developed a stochastic model for PCR systems.

The left model in figure 2 describes the enzyme states, while the right one describes the substance states:

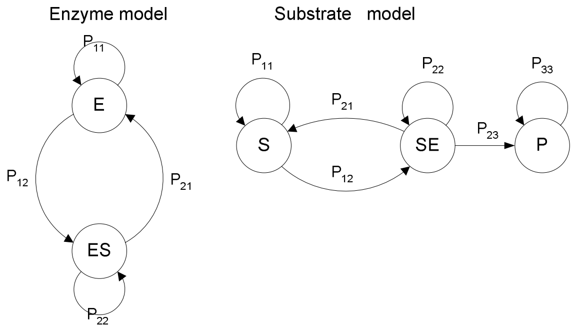

Figure 2: Enzyme-Substance Markov models

Transitions between the component states are noted by arcs, and each arc is associated with a transition probability alone the arc. In a general Markov model, the transactions between the various states from time instant n-1 to time instant n, are given by:

$$\pi^{n} = P^{T} \pi^{n-1} \qquad (2)$$

where the matrix P represents the transition probabilities between the various states:

$$\{P\}_{ij} = P_{ij} \qquad (3)$$

P_{ij} is the probability to move from state i to state j, and π is the vector probabilities to be in each state.

$$\pi = \begin{bmatrix} \pi_1 \\ \pi_2 \\ \vdots \end{bmatrix} \qquad (4)$$

For the process of equation (1), the enzyme Markov model equation is:

$$\begin{bmatrix} \pi_E \\ \pi_{ES} \end{bmatrix}^{n+1} = \begin{bmatrix} P_{11} & P_{21} \\ P_{12} & P_{22} \end{bmatrix}^{n} \cdot \begin{bmatrix} \pi_E \\ \pi_{ES} \end{bmatrix}^{n} \qquad (5)$$

The substrate Markov model equation is:

$$\begin{bmatrix} \pi_S \\ \pi_{SE} \\ \pi_P \end{bmatrix}^{n+1} = \begin{bmatrix} P_{11} & P_{21} & 0 \\ P_{12} & P_{22} & 0 \\ 0 & P_{23} & 1 \end{bmatrix}^{n} \cdot \begin{bmatrix} \pi_S \\ \pi_{SE} \\ \pi_P \end{bmatrix}^{n} \qquad (6)$$

The different states probabilities are defined as follows:

$$\pi_E^{(n)} = \frac{E[n]}{E[0]} \quad (7) \qquad \pi_{ES}^{(n)} = \frac{ES[n]}{E[0]} \quad (8)$$

$$\pi_{ES}^{(n)} = \frac{ES[n]}{E[0]} \quad (9) \qquad \pi_S^{(n)} = \frac{S[n]}{S[0]} \quad (10)$$

$$\pi_{SE}^{(n)} = \frac{ES[n]}{S[0]} \quad (11) \qquad \pi_P^{(n)} = \frac{P[n]}{S[0]} \quad (12)$$

Where S[n] denotes the quantity of the material S at instant n. The assumptions used in order to simplify the model equations can be divides into three categories:

1. Biochemical: steady state approximation.
2. Statistical: exponential behavior of enzymes.
3. Conceptual: probabilities-concentrations duality.

Under these assumptions, steady state probabilities P_{ij} exist and are given by[3]:

$$p_{12} = k_1 \cdot [S] \cdot \Delta \quad (13)\,; \quad p_{21} = (k_2 + k_{-1}) \cdot \Delta \quad (14)$$

where Δ is the observation period.

The resulting steady state probability matrices for the two Markov processes are:

$$P_{enzyme} = \begin{bmatrix} 1 - k_1 \cdot \Delta \cdot [S] & (k_{-1} + k_2) \cdot \Delta \\ k_1 \cdot \Delta \cdot [S] & 1 - (k_{-1} + k_2) \cdot \Delta \end{bmatrix} (15)$$

$$P_{Subs\tan ce} = \begin{bmatrix} 1 - \Delta \cdot k_1 \cdot [E] & \Delta \cdot k_{-1} & 0 \\ \Delta \cdot k_1 \cdot [E] & 1 - \Delta \cdot (k_2 + k_{-1}) & 0 \\ 0 & \Delta \cdot k_2 & 1 \end{bmatrix} (16)$$

The transition probabilities are expressed in terms of physical terms, which can be measured for any given enzymatic process.

4. ILLUSTRATION

We demonstrate our approach by presenting the detailed state diagram the NER repair process of the *E.coli* bacteria[7], a multi-step system of the following stages:

1. A continuous scan of the genome in a search for induced damage in the DNA.
2. A damage detection and verification mechanism.
3. Pre DNA double helix processing (the unwinding of the double helix).
4. A three stage process aimed to the removal of the damaged oligonucleotide.
5. The initialization and the recruitment of the cell synthesis machinery, which results in a synthesis and sealing of a newly DNA strand.

The building blocks of the NER repair path are a group of six enzymes named: uvrA, uvrB, uvrC, uvrD, DNA polymerase and DNA ligase

The NER operation starts with the formation of an enzyme complex composed of two uvrA enzymes and a single uvrB enzyme, denoted by uvrA_2B. The uvrA_2B complex scans the DNA and searches for damages in it; it is capable of binding to the DNA (both to damaged or undamaged DNA, but yet with different probabilities). After the complex mounts the DNA, it begins to move along it, while searching for damage. On encountering a suspicious area, it initiates a series of operations purposed to repair the damage. In this process the uvrA enzyme serves as a molecule matchmaker. We describe its operation by identifying the following states (see Fig. 3). A single enzyme molecule (A) is the initial state. It can form an uvrA dimer (A_2) with probability P_{12}. The dimer can dissociate spontaneously (with probability P_{21}) or can form an uvrA_2B enzyme complex (A_2B) with probability P_{23}. Once an uvrA_2B complex has been formed, it begins its scanning operation climbing the DNA at an arbitrary location and starting its translocation along it, while searching for damages.

The uvrA_2B complex has a probability P_{34} to climb on the DNA at non damaged location, and probability P_{37} to bind directly to a damaged location. The translocation phase is modeled by two states: tA_2B+DNA(n) and tA_2B+DNA(n+1), representing uvrA_2B pre- and post-translocation activity along the N nucleotides DNA molecule. The translocation activity is modeled by P_{45}, P_{54} and P_{53}, P_{63}, which represent successive translocation verse random dissociation of uvrA_2B from the DNA.

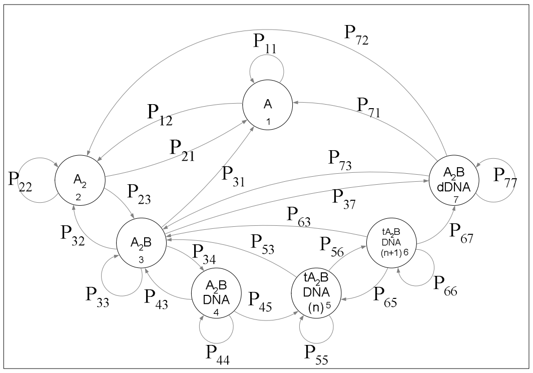

Figure 3: **The Markov state diagram for the uvrA enzyme**

Once damage has been reached, the uvrA_2B enzyme halts (A_2B+dDNA) and verifies the damage. As the operation of the uvrA_2B enzyme ends the DNA is bent by 135 degrees and the uvrB enzyme stays attached to the damage.

Assuming a steady state conditions, the probabilities to be in each of the states can be evaluated by solving:

$$\pi = P^T \cdot \pi \ \rightarrow \left(P^T - I\right) \cdot \pi = 0 \qquad (17)$$

Similarly, the steady state probabilities of each of the enzymatic processes can be derived given the transition probability matrix P. The individual transition probability P_{ij} between states i and j can be derived from the enzyme kinetic equation and is a function of the enzymes, substrate concentrations and the process rate velocities [5].

5. DISSCUSION

The DNA repair sub-systems may have mistakes during the different repair stages: detection, verification phase, or even at the synthesis stage. Such mistakes lead to errors in the results of the DNA repair process. We suggest a way to quantify these errors in probabilistic term.

Due to the complexity of the processes, scientists usually relate certain diseases with bio-molecular parameters based on cumulative experience. For the first time, our model will enable systematic analysis of the statistical performance, as a function of bio-chemical parameters as enzymes concentration and rate velocities, for each of the DNA repair mechanisms, in order to quantify them and to understand how various changes in local bio-molecular level influence the resulting error.

Knowing the enzymes' statistical model and the interactions between the various enzymes also provides a useful tool to analyze the system's overall performances, allowing to determine the probabilities of some basic events in the system, such as the probability for miss detection, or miss damage repair.

In the post-genomic era, gene expression of the different components of the DNA repair processes are being heavily investigated. Our novel approach can be used to relate identified genes or mutations to the probability of having certain DNA damage. Moreover, understanding the DNA repair mechanism may also contribute to the development of new approach to error-correcting coding in technology (e.g., communication and computer systems).

6. REFERENCES

1. B. A. Sokhansanj, G. R. Rodigue, J. P. Fitch and D. M. Wilson III, "A quantative model of human DNA Base excision repair mechanistic insights", *Nucleic Acids Research* 2002, Vol.30, No 8 1817-1825.
2. J. Ricard, "What do we mean by biology complexity?" *C.R Biologies*, 326 (2003) 133-140.
3. R. Sever, ***System approach to the DNA Repair Process in*** *E.coli,* M.Sc Thesis, Tel Aviv University, April 2004.
4. T. M.Cover J. A. Thomas, ***Element of Information Theory***, John Wiley, 1991.
5. J. A. Nickoloff, M. F. Hoekstra, ***DNA damage and repair***, Totowa N.J: Humana Press 1998
6. A. Hassibi, H. Kakavand and T. H.Lee*, "A stochastic model and simulation algorithm for polymerase chain reaction (PCR) system",* In Proceedings of GENSIPS'2004.
7. Y. Zou, C. Luo, and N. E. Geacintov. "*Hierarchy of DNA Damage Recognition in Escherichia coli Nucleotide Excision Repair", Journal Biochemistry*, 40 (9), 2923 - 2931, 2001.