OPTIMIZATION CRITERIA FOR DNA REPAIR

Ronit Bustin and Hagit Messer

School of Electrical Engineering, Tel Aviv University, Tel Aviv 69978, ISRAEL

ABSTRACT

Engineers design error-correction systems to try to achieve zero error (that is, a perfect matching between the transmitted and the received code) while error-free transfer of genetic information from one generation to the other will stop evolution. That is, in biological systems, unlike the case of person-made systems, the target of error-correction procedures is to maintain a certain level of errors – not too high (to ensure the sustainability of the individual) and not too low – to ensure the sustainability of life. We suggest a two-part model for the communication of the genetic information from one generation to the other. The DNA repair process plays different role in each part of this model. In this paper we will focus only on the first part of the suggested model. We suggest RNA polymerase (RNAp) as the optimization criterion for the efficiency of this first part, that is, the "success" of the DNA repair process is a successful operation of the RNAp in spite of a lesion.

1. BACKGROUND & MOTIVATION

In this paper we study the DNA repair process from an engineering point of view, where we look at the genetic information flow from one generation to another as a communication system. 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. Previous attempts to model this system were incomplete and cannot be used for system analysis and/or simulation. Based on the modeling of the components of the DNA repair system by a probability Markov state diagram [16,17,18], we show[1] that better modeling of the overall system can be acomplished. Our statistical signal processing approach to DNA repair is different from the one taken previously, by e.g., John Hopfield [8], who developed an error correcting model known as "kinetic proofreading". While others have focused on proofreading in the individual process, we aim to understand how the different kinetic parameters of each enzyme influence the cell overall capability to deal with DNA damages.

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

1. Nucleotide excision repair (NER).
2. Base excision repair (BER).
3. Mismatch repair (MMR).
4. DNA repair by damage reverse (DR).
5. Recombination repair.
6. Translation DNA synthesis (TLS).

The interaction between the different repair mechanisms from the fuctional point of view needs to be explored in order to build a comprehensive model to relate the initial condition (damaged DNA) and the final states (corrected DNA, or a mutation).

In this paper we suggest a simplified model for the cell's genetic information flow from generation n to generation n+1. also In particular, we suggest a novel optimization criterion for the first module of the genetic communication system which is based on its functional setting.

The paper is organized as follows: section 2 presents a conceptual model of the genetic information flow from one generation to the other as a communication system. Note that the conceptual model consists of two parts, where the rest of the paper is devoted to the first part only. Section 3 suggests a novel optimization criterion for the first part of the model, and the relevant biological information presented from the lesion point of view. Section 4 concludes with a discussion on open questions and future directions.

2. THE CELL'S GENETIC INFORMATION FLOW MODEL

The main task in modern communication is reliably transmitting and receiving information between different points on a given physical media. Living cells face task similar to

email:messer,bustin@eng.tau.ac.il

[1] An approach similar to ours has been taken in the work of Hassibi et. al [7] which developed a stochastic model for PCR systems.

communication systems – they need to successfully deliver the genetic information from one generation to the other. However, in order to survive, at least to the reproduction point, they must continuously deal with both external and internal hazards which threat to damage the integrity of the cell's DNA. A "success" in terms of molecular biology may be considered as the cell ability to reproduce, and to pass its genetic material to its offspring. We have suggested [18] to model the cell's genetic information flow from generation n to generation n+1 by the simplified model depicted in Figure 1.

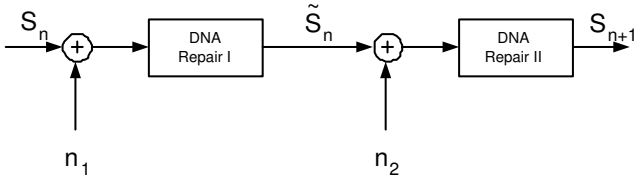

Fig. 1. Schematic genetic information flow in a cell. S_n is the DNA of the cell at generation n; $\widetilde{S}_n$ is the DNA pre the replication process; S_{n+1} is the DNA at generation: n+1.

The input signal S_n represents the cell 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 $\widetilde{S}_n$ represents the DNA just before the replication. In general, $S_n \neq \widetilde{S}_n$ because during the cell's life, damage is introduced to the DNA (represented by the additive noise n_1). The damage is partially corrected by the first DNA repair module (denoted by "DNA repair I" in the block Diagram of Figure 1). Replication errors are represented by the additive noise process n_2 and are partially corrected by the second DNA repair module.

This model suggests putting the various DNA repair mechanisms into two groups. The first is a general repair process, which is composed of repair mechanisms as the BER, the NER and the DR. These mechanisms operate along the cell life cycle (both modules of Figure 1) and are responsible to correct any damage to the DNA, both internal and external lesions. The second group includes the mismatch repair mechanism, which is responsible to correct DNA polymerase mismatches, and recombination repair, TLS. Both are responsible to overcome replication blocks and enable completion of DNA replication (the second module of Figure 1). The two modules 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.

In communication systems, a common tool for measuring the difference between two code words is the Hamming distance:

$$D(s_1, s_2) = \sum_i (s_1(i) \neq s_2(i)) \qquad (1)$$

This measure weights errors in different bits along the code uniformly. The ability to model the genetic information flow as a communication system suggests on using similar performance measures as the Hamming distance. Indeed, so far, the biological evidence points at a more generic kind of repair mechanism that, in signal processing terms, "acts on individual bits", and is more dependent on the geometry and type of damage rather than the DNA sequence content. However, using the Hamming distance as a measure of "success" for the DNA repair process is artificial, since the probability of repairing errors in different parts of the genome is not the same. Errors in different parts of the genome have different meaning and may not be weighted uniformly.

Moreover, while engineers design error-correction systems to try to achieve zero error (that is, a perfect matching between the transmitted and the received code), error-free transfer of genetic information from one generation to the other will stop evolution. That is, in biological systems (unlike the case of person-made systems), the target of error-correction procedures is to maintain a certain level of errors – not too high (to ensure the sustainability of the individual) and not too low – to ensure the sustainability of the species [13,11,15].

In the two-steps model of Figure 1 we separate the DNA repair process into two modules, suggesting, for each one, a different performance measure. Our approach is to establish a performance measure from the point of view of the main function of the DNA repair block. The different functions of the two blocks can be summarized as follows: The first part of the process of Figure 1 aims to enable the cell to live till reproduction, while the second part is the one in which the reproduction process actually transfers the genetic information to the next generation. In this paper we concentrate on the first module of Figure 1, and we suggest an appropriate optimization criterion for its operation.

3. RNA POLYMERASE AS AN OPTIMIZATION FUNCTION

The survival of the individual cell requires the stability of its cell functions. Changes to its proteins can result in mal functioning and even a death of the cell. This leads us to the hypothesis that the goal of the individual is to maintain stability in its proteins, that is – maintaining the genetic information used by the individual in the form of transcription

to messenger RNA (mRNA) [1]. Transcription to mRNA is a mechanism that operates along the cell life cycle, until it reaches the replication stage, that is, during the first DNA repair module in Figure 1. We propose to evaluate the performance of the "DNA repair I" mechanisms, from the point of view of RNA polymerase (RNAp), the main player in transcription operation. That is, we suggest that a "success" of the "DNA repair I" process equals to a successful operation of the RNAp in spite of the lesion.

RNAp reacts in a different manner to different lesions of the DNA. Not in all cases the consequences are mutations in mRNA, and not in all cases the mutations in the mRNA are lethal. We suggest to set the RNAp as our optimization criterion for the first module of Figure 1, that is – the aim of the first module repair mechanisms of the DNA is to minimize the lethal mutations in mRNA.

Moreover, we can scale the damage level a specific mRNA outcome may have on the cells ability to survive. One can look on the following set of possible mRNA outcomes:

1. correct mRNA: the mRNA sequence if flawless.
2. silent mutation in the mRNA: most amino acids are encoded by several different codons. Silent mutations are in a nucleotide that alters a codon without changing the encoded amino acid.
3. blockage: a mutation to the DNA that blocks transcription, that is, no creation of mRNA.
4. missense mutation in the mRNA: the new nucleotide alters the codon so as to produce an altered amino acid in the protein product.
5. nonsense mutation in the mRNA: the new nucleotide changes a codon that specified an amino acid to one of the STOP codons. Therefore, translation of the mRNA transcribed from this mutant gene will stop prematurely. The earlier in the gene that this occurs, the more truncated the protein product and the more likely that it will be unable to function.

The effect of the last two mRNA outcomes depends also on the effect the mutation will have on the created protein, that is, it depends on the "reading frame" of the specific protein translated from the mRNA.

Consider a mechanism whose input is a specific level of damage to the DNA, caused by a specific damaging agent such as UV light, oxidation, alkylation etc. $\widetilde{S}_n(i)$ is a small segment of $\widetilde{S}_n$ which, under a specific damaging agent, and a specific probability for damage, can be damaged only once. We will mark the transcriptom of $\widetilde{S}_n(i)$ as $\widetilde{\widetilde{S}_n(i)}$ and in a similar manner, the transcriptom of $S_n(i)$ as $\widetilde{S_n(i)}$. Under these notation, our hypothesis equals to the following:

$$min(\sum_i Pr(\widetilde{\widetilde{S}_n(i)} \neq \widetilde{S_n(i)}) \cdot W_i) \tag{2}$$

Where i goes over all segments of the DNA coding to mRNA, and W_i denotes the weight that distinguishes between the impact of the mutation, lethal mutations at one end, and silent at the other. By quantizing each mutation to one of the mentioned possible mRNA outcomes, equation (2) can be simplified to :

$$\begin{aligned} min(\sum_i (& \\ & Pr(\widetilde{\widetilde{S}_n(i)} \in \text{correct mRNA}) \\ & \cdot W_{\text{correct}} + \\ & Pr(\widetilde{\widetilde{S}_n(i)} \in \text{no mRNA would be created}) \\ & \cdot W_{\text{no mRNA}} + \\ & Pr(\widetilde{\widetilde{S}_n(i)} \in \text{mRNA with silent mutation}) \\ & \cdot W_{\text{silent}} + \\ & Pr(\widetilde{\widetilde{S}_n(i)} \in \text{mRNA with missense mutation }) \\ & \cdot W_{\text{missense}} + \\ & Pr(\widetilde{\widetilde{S}_n(i)} \in \text{mRNA with nonsense mutation}) \\ & \cdot W_{\text{nonsense}} \\ &)) \end{aligned} \tag{3}$$

Where there are five possible weights, $W_i, i = 1, \ldots, 5$, for each of the possible mRNA outcomes, instead of a different weight for each lesion of the transcriptom segment.

To validate our hypothesis we need to calculate the different probabilities for each of the possible mRNA outcomes. For each lesion in the DNA there are several repair mechanisms that can fix it. This fact leaves us with numerous repair probabilities, that is, different repair probabilities for almost each and every lesion. For this reason we decided to examine the DNA repair mechanisms from the lesions point of view. This approach, together with the novel suggested optimization criterion, will also suggest on the interconnection between the different DNA repair mechanisms.

Our study of the repair mechanisms from the lesion point of view, yields the following questions: What harms the DNA and creates the lesion? What effect does the lesion have on the repair enzymes? What are the possible repair outcomes? How does RNAp react to the lesion? In order to answer these questions we have summarized the related biological information [21,23,12,10,6,26,3,9,2,24,25,5,19,20, 4,22,14] in table 1.

Table 1. Damages inflicted upon the DNA and the transcription mechanism reaction to these lesions.

Damaging agent	*lesion*	*BER*	*NER*	*TCR*	*Damage* Reverse	*RNAp: block,pass efficiency*	*RNAp Mutagencicity*	*Effect on Protein*
oxidation	8-oxoguanine	+	+	+	−	bypass high eff.	A 50% C 50%	A instead of C 37% missense
UV light, pollutants	AP site	+	+	?	−	bypass moderate eff.	A 75% deletion ∼ 15%	A instead of C 74% missense
hydrolytic deamination	uracil	+	−	−	−	bypass high eff.	primeraly A	A instead of G 68% missense
UV light	CPD	+	+	+	+	blocks high eff.	−	−
alkylation	O_6 - methylguanine	−	+	−	−	bypass high eff.	primeraly U	U instead of C 62.5% missense
deamination & oxidation	5,6-Dihydrouracil	+	−	−	−	bypass high eff. (brief pausing)	primeraly A	A instead of G 63% missense
UV light	6-4 photo.	−	+	+	−	blocks high eff.	−	−
oxidation	TG − thymine glycol	+	+	+	−	bypass 50% (of the time)	?	?
alkylation	7-Methylguanine	+	−	−	−	bypass high eff.	?	?
Psoralen and UV	HMT	−	+	−	−	blocks high eff.	−	−
	SSB − single strand breaks	−	−	?	−	bypass variable eff.	deletion of exact gap size	frameshift
alkylation	3-Methyladenine	+	?	−	−	bypass moderate	?	?

Table 1 describes several main lesions (column 2) and it includes both the biological information regarding the repair pathways of these lesions (columns 3 to 6), and the reaction of the transcription mechanism to the lesion (the last three columns). The reaction of the transcription mechanism includes two main aspects: the reaction of RNAp to the lesion, and the effect of such a reaction on the final protein output.

The information in table 1 allows us to build a Markov diagram for each lesion, describing the different possible repair pathways of that lesion. Given the additional biochemical information: initial concentrations, diffusion coefficients, rate laws and rate constants, will allow us to calculate the probabilities of the different mRNA outcomes. An example for such a Markov diagram, for the case of Cyclobutane Pyrimidine Dimers (CPDs) inflicted upon the DNA sequence as a result of UV light, is given in Figure 2[m1].

4. SUMMARY

One may wonder about the appropriate way to weight the different events. Is “no mRNA creation” more crucial for the cells survival than “missense mutation in the mRNA”? Is “silent mutation in the mRNA” crucial in any way? Our initial assumption is that low probability of a certain outcome suggests a higher threat on the survival of the cell. Therefore, higher probability for repairing it by the cell’s repair mechanisms is anticipated, trying to avoid such outcomes. Our goal is to test this assumption in addition to our primary assumption, that RNAp is the optimization criterion for the efficiency of the first module of the DNA repair. In order to do so, we need to calculate the probabilities of mRNA outcomes under different damaging agents. Viewing similar relations between the probabilities of the different mRNA outcomes for different damaging agents, will suggest an optimization according to RNAp. That is, if for most damaging agents a specific mRNA outcome has relatively low probability, suggests that this specific outcome strongly threats the survival of the cell. Such a conclusion will validate our initial assumption.

Acknowledgment

The authors would like to thank Prof. Zvi Livneh from the Weizmann institute and Mr. Ram Sever for our fruitful discussions.

5. REFERENCES

[1] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, P. Walter, Molecular Biology of the Cell, Taylor & Francis, fourth edition, 2004.

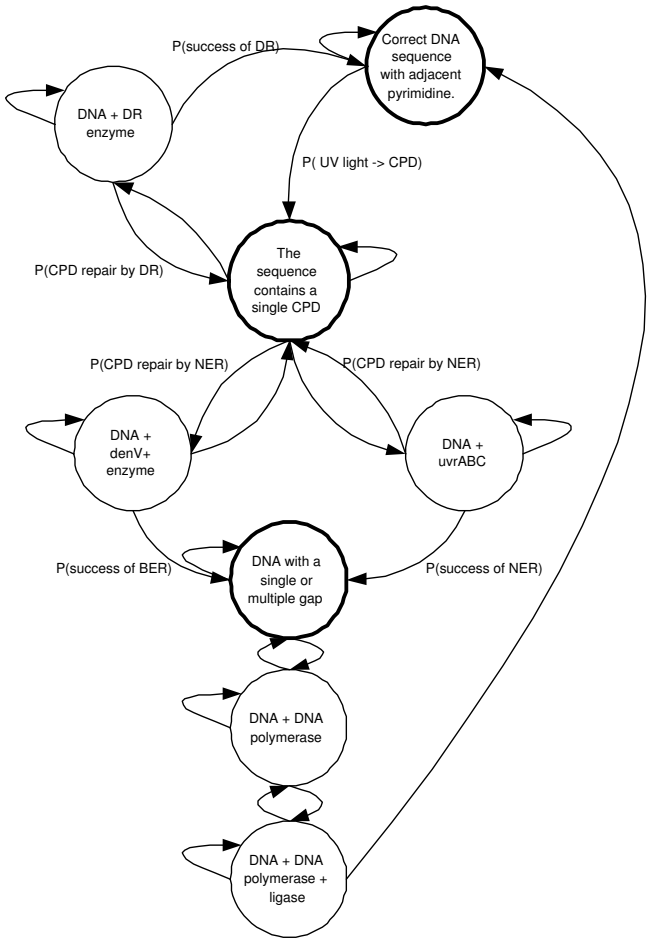

Fig. 2. The Markov diagram for Cyclobutane Pyrimidine Dimers (CPDs) inflicted upon the DNA sequence as a result of UV light.

[2] Nasser R. Asad, Lidia M.B.O. Asad, Carlos E.B. de Almedia, Israel Felzenszwalb, Januario B. Cabral-Neto & Alvaro C. Leitao, Several pathways of hydrogen peroxide action that damage the *E. coli* genom, Genetics and Molecular Biology, Vol. 27, No. 2 291-303, 2004.

[3] M.S. Cooke, M.D. Evans, M. Dizdarogulu & J. Lunec, Oxidative DNA damage: mechanisms, mutation, and disease, The FASEB Journal, Vol. 17, 1195-1214, July 2003

[4] P.W. Doetsch, Translation synthesis by RNA polymerases: occurrence and biological implications for transcriptional mutagenesis, Mutat. Res. 510:131-140 (2002).

[5] B.A. Donahue, S. Yin, J.S. Taylor, D. Reines & P.C. Hanawalt, Transcription cleavage by RNA polymerase II arrested by a cyclobutane pyrimidine dimer in the DNA template, Proc. Natl. Acad. Sci. USA, Vol 91, pp. 8502-8506, August 1994.

[6] W. Franklin & W.A. Haseltime, Removel of UV light-induced pyrimidine-pyrimidone(6-4) products from *Escherichia coli* DNA requires the uvrA, uvrB, & uvrC gene product, Pro. Natl. Acad, Sci. USA Vol. 81, pp. 3821-3824, June 1984.

[7] 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

[8] J.J Hopfield, Kinetic Proofreading: A New Mechanism for Reducing Errors in Biosynthetic Processes Requiring High Specificity, Proc Natl Sci October 1974.

[9] Yun-Song Lee & Myung-Hee Chung, Base excision repair synthesis of DNA containing 8-oxoguanine in *Escherichia coli*, Experimental and mulecular medicine, Vol. 35, No. 2, 106-112, April 2003

[10] J.J. Lin, A. Sancar, A new mechanism for repairing oxidative damage to DNA: ABC excinuclease removes AP sites and thymine glycols from DNA. Biochemistry 28:7979-7984 (1989).

[11] E.E. May, Comparative Analysis of Information Based models for Initiating Protein Translation in *Escherichia coli* K-12, M. S. thesis, NCSU, December 1998.

[12] Jac A. Nickoloff, Merl F. Hoekstra, DNA Damage and Repair, Totowa N.J: Humana Press 1998

[13] S. Noorbaloochi and A. H. Tewfik, Overview of DNA Demage Detection, preprint, 2004.

[14] Jacques Ricard, What do we mean by biology complexity?, C.R Biologies, 326 (2003) 133-140.

[15] G. Rosen and J. Moore, Investigation of Coding Structure in DNA. IEEE International Conference on Acoustics, Speech, and Signal Processing Proceedings, Vol. 2, pp. 361-364, April 2003.

[16] R. Sever, A statistical model for DNA repair mechanisms, thesis under the supervision of Prof. Hagit Messer, April 2004.

[17] R. Sever, H. Messer, System approach to the DNA Repair Process in *E.coli*, In Proceedings of GENSIPS'2004.

[18] R. Sever, H. Messer, On the use of Sequential Monte Carlo to Approximate the Optimal Filter Sensitivity for Parameter Estimation, ICASSP 2005.

[19] D.A. Scicchitano, I. Mellon, Transcription and DNA damage: a link to a kink. Environ. Health Perspect. 105(Supplement 1):145-153 (1997).

[20] D.A. Scicchitano, E.C. Olesnicky, A. Dimitri, Transcription and DNA adducts: what happens when the message gets cut off? DNA repair 3: 1537-1548 (2004).

[21] A. Snowden, Y.W. Kow, B. Van Houten, Damage repertoire of the *Escherichia coli* UvrABC nuclease complex includes abasic sites, base-damage analogues, and lesions containing adjacent 5' or 3' nicks. Biochemistry 7, 29(31):7251-9 (1990).

[22] S. Tronaletti, P.C. Hanawalt, Effect of DNA lesions on transcription elongation. Biochimie 81:139-146 (1999).

[23] B. Van Houten, Nucleotide excision repair in *Escherichia coli*. Microbiol. Rev. 54:18-51 (1990).

[24] A. Viswanathan and P. W. Doetsch, Effects of Nonbulky DNA Base Damages on *Escherichia coli* RNA Polymerase-mediated Elongation and Promoter Clearance. J. Biol. Chem., 273(33): 21276-21281 (1998).

[25] A. Viswanathan, J. Liu, P.W. Doetsch, *E. coli* RNA Polymerase Bypass of DNA Base Damage-Mutagenesis at the Level of Transcription. In: Molecular Strategies in Biological Evolution (L.H. Caporale, ed.) Annals, 870:386-388. The New York Academy of Sciences, New York (1999).

[26] Yue Zou, Charlie Luo, and Nicholas E. Geacintov, Hierarchy of DNA Damage Recognition in *Escherichia coli* Nucleotide Excision Repair, Journal Biochemistry, 40 (9), 2923 -2931, 2001.