Information-Theoretic Model of Evolution over Protein Communication Channel

Liuling Gong, *Student Member, IEEE*, Nidhal Bouaynaya, *Member, IEEE*, and Dan Schonfeld, *Senior Member, IEEE*

***Abstract*—In this paper, we propose a communication model of evolution and investigate its information-theoretic bounds. The process of evolution is modeled as the retransmission of information over a protein communication channel, where the transmitted message is the organism's proteome encoded in the DNA. We compute the capacity and the rate-distortion functions of the protein communication system for the three domains of life: Archaea, Bacteria and Eukaryotes. The tradeoff between the transmission rate and the distortion in noisy protein communication channels is analyzed. As expected, comparison between the optimal transmission rate and the channel capacity indicates that the biological fidelity does not reach the Shannon optimal distortion. However, the relationship between the channel capacity and rate distortion achieved for different biological domains provides tremendous insight into the dynamics of the evolutionary processes of the three domains of life. We rely on these results to provide a model of genome sequence evolution based on the two major evolutionary driving forces: mutations and unequal crossovers.**

***Index Terms*—Protein communication system; Channel capacity; Rate-distortion theory; Non-homogeneous Poisson process.**

I. Introduction

In this work, we describe the evolutionary process of transmitting information from generation to generation using communication and information theory. The process of transmission of genetic material during reproduction resembles the engineering system of transmission of information over a channel. Every organism contains the DNA, or the genome sequence, which encodes the information required to create proteins, the functional machinery of the organism. During cell duplication or reproduction, the genomic material is copied to create the offspring's genome. This duplication of genetic material is typically error-prone [1]. By decoding the genome into proteins, the organism come into being. The decoding process is almost universal for all organisms and is called translation in molecular biology. Hence, we have a biological system, which is composed of three elements: the encoded message (DNA), a noisy medium of transmission or channel (DNA storage and replication), and a decoder (the translation process). Since the output of the decoder is the organism's proteome, and the objective of a communication system is to receive messages from a source and to transmit them through a channel to a destination (see Fig. 1), the source of the biological communication system should generate the proteome. Forthwith, we observe that there are two main differences between the biological information processing system and the communication engineer system: The first is that biology does not encode proteins into DNA. It only decodes genes into proteins. The second is that, unlike the communication engineer system, the biological communication system is not designed to minimize transmission errors. Otherwise, evolution will not be possible. Intuitively, there has to be a balance between keeping the cell identity by reliable transmission of its protein set and allowing errors to occur purposefully to encourage evolution.

The biological communication system is shown in Fig. 2, and we will refer to it as the protein communication system [2] since the transmitted and received messages are protein sequences. It is important to reiterate that the encoding process, in the protein communication system, is only a mathematical model of the protein information captured by the DNA. In order to clarify this abstraction, let us use the following analogy with an engineering communication system for video transmission: We want to transmit a video stored in a computer to other computers. The initial computer maintains an MPEG code of the video. Assuming that the computer at the receiver has the decoder required to decode MPEG files into videos, transmission of the video message to other computers only requires sending the corresponding MPEG code. At the receiver, the MPEG file will be decoded to display the desired video. Assume further that the first MPEG code was created by chance. Therefore, this system never encodes a video into MPEG. It only decodes MPEG to display a video. Nonetheless, the proper communication model for this video transmission system relates to the transmission of the video between the sender and receiver. Note that, in this system, the only signal transmitted is the MPEG code and not the video. We also note that, although the MPEG file is decoded by the receiver to reconstruct the video, the original video was never encoded by the sender. Yet, from an engineering communication system perspective, the information transmitted between the sender and receiver relates to the video, whereas the MPEG code is simply used to represent the video over the communication channel; i.e. "video $\rightarrow$ MPEG $\rightarrow$ MPEG $\rightarrow$ video" even though the process "video $\rightarrow$ MPEG" never takes place. Figure 3 summarizes the analogy between this engineering video communication system and the protein communication system. The video is analogous to the proteome of the cell; the MPEG file is analogous to the genome or the DNA sequence. The encoding process (from protein to DNA) is bypassed in Nature by ensuring that organisms maintain both proteins and DNA. The errors introduced in the protein communication channel, during transmission, correspond to the errors introduced during

Digital Object Indentifier 10.1109/TCBB.2009.1

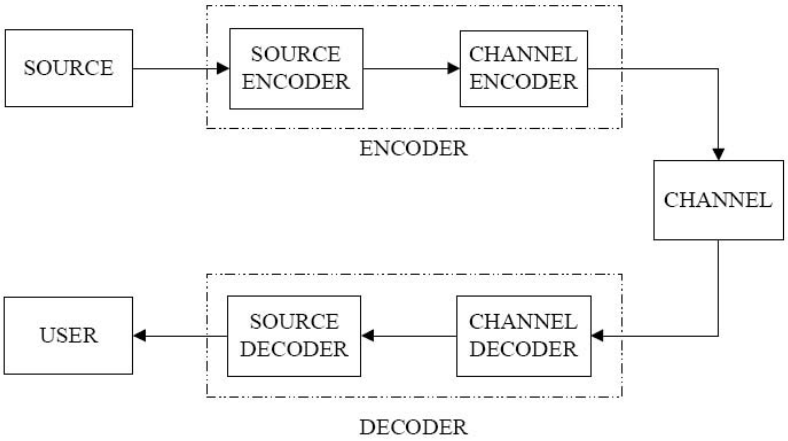


Fig. 1. A communication system block diagram

storage and replication of the encoded message (i.e., DNA).

A protein communication model which models the transmission of information in sexual reproduction, is much more mathematically involved than the single source communication system in cell replication or asexual reproduction. Analysis of such communication system requires the knowledge of multi-user information theory and distributed coding, and will not be discussed in this paper. However, as a first order approximation, the analysis of the information-theoretic bounds of the single source protein communication system will unveil a great deal about the optimality of biological systems from an information-theoretic perspective. Using the mathematical model of protein communication system, we will translate the problem of a species' evolution into the language of mathematics, in particular the language of communication theory. The problem of a species' evolution will be represented as the iteration of a communication channel over time.

The study of information theory begun with the revolutionary work of Claude Shannon in the early years of World War II [3]. One of Shannon's great contributions to the field of information theory is the separation of the semantic content of a message from the dynamic channel that transmits the message. A particular information source may be, for example in the context of the Internet, an audio file, a video clip, or an email. In the context of the protein communication channel, the information source can be the proteome of a bacterium, a person, or any organism. However, the design of a communication channel is not for a particular message or type of message; rather, the transmission machinery is designed for all possible messages, regardless of their semantic meanings. This explains, from an information-theoretic viewpoint, why the biological information storage and transmission system is common (with rare variations) to all living organisms. In this context, the transmission of genetic information can be investigated in the perspective of information theory. In particular, it is legitimate to ask at what rate can the genomic information be transmitted. And what is the average distortion between the transmitted message and the received message at this rate? Shannon's noisy-channel coding theorem states that, by properly encoding the source, a communication system can transmit information at a rate that is as close to the channel capacity as one desires with an arbitrarily small transmission error [4]. Conversely, it is not possible to reliably transmit at a rate greater than the channel capacity. However, the theorem is not constructive and does not provide any help in designing such codes. Nonetheless, in the case of biological communication systems, evolution has already designed the code, such that the encoded message is the DNA sequence. Comparison of the genomic transmission rate with the channel capacity will reveal whether the genomic code is efficient from an information-theoretic perspective. However, even if the channel capacity is not exceeded, we are assured that biological communication systems do not rely on codes that produce negligible errors since the level of distortion presented must account for evolutionary processes. Therefore, it is interesting to ask ourselves whether biological communication systems maintain an optimal balance between the transmission rate and the desired distortion level needed to support adaptive evolution. Rate-distortion theory analyzes the optimal tradeoff between the transmission rate, $R(D)$, and distortion, D, in noisy communication channels [5]. The theory is used to characterize the minimal rate required for transmission of information over a channel in order for the receiver to be able to reconstruct the transmitted message without exceeding a given distortion [4] [5]. Given the fidelity, D, which is presented in biological communication systems, we can compare the genomic transmission rate with the optimal rate $R(D)$ to determine whether the genomic code achieves the optimal rate-distortion criterion. Moreover, by equating the optimal rate $R(D)$ with the channel capacity, C, we can determine whether the biological fidelity, D, reaches the Shannon optimum distortion.

In [6], we experimentally computed the channel capacity of the protein communication system, and the rate-distortion curves of the three branches of life: Archaea, Bacteria and Eukaryotes. We found that for low-distortion regions, Prokaryotes are more efficient than Eukaryotes, in the sense that they correspond to a lower distortion for a given transmission rate, whereas Eukaryotes are more efficient than Prokaryotes for higher distortion regions. In this paper, we elaborate upon this

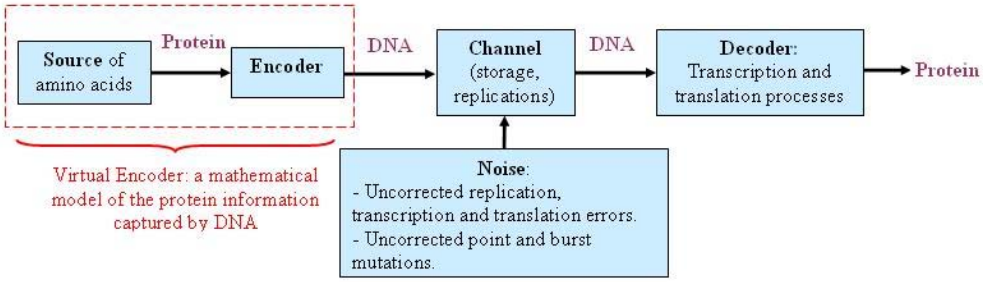


Fig. 2. Protein communication System. The protein communication model is isomorphic to the engineering communication system: It is composed of an encoder, an error-prone channel and a decoder. The encoder is only a mathematical model of the protein information captured by the DNA. Nature was able to bypass the encoding process by ensuring that organisms maintain both DNA and proteins. Furthermore, based on the highly redundant structure of the DNA sequence (e.g., presence of a large percentage of non-coding segments), we argue that the encoder models a source and channel encoder. The physical channel models the transmission and storage of the DNA and is the source of errors. The encoded DNA sequence is transcribed into mRNA; then decoded by the ribosomes from the 4-letter alphabet mRNA sequence to the 20-letter alphabet amino-acid chain in the protein. The decoding process, called translation in molecular biology, is accomplished based on the well-known genetic code.

Engineering Communication System	Protein Communication System
Video	Set of proteins of the cell
MPEG	DNA
Encoder	—
Decoder	Translation Process
No Error Desired	Some Errors Desired

Fig. 3. Comparison between the engineering communication system for video transmission and the protein communication system during cell replication.

analysis and expand it by providing an information-theoretic model of evolution for Eukaryotes and Prokaryotes based on the two major evolutionary processes: mutations and unequal-crossover. The proposed model explains and validates the experimental rate-distortion curves observed in [6].

The rest of this paper is organized as follows: In Section II, we briefly review the protein communication system introduced in [2]. In Section III, we compute the protein channel capacity. We show that organisms with lower mutation rates have higher channel capacity and therefore can transmit their genetic information reliably at a higher rate. In Section IV, we propose an evolutionary model based on mutations and unequal crossovers and then compute the average distortion in Prokaryotes and Eukaryotes. We subsequently demonstrate that the actual rate distortion curves of Eukaryotes, Bacteria and Archaea are in accordance with the proposed model. Finally, in Section V, we highlight the implications of the proposed evolutionary model in various aspects of biology, computational biology and genetic engineering, especially in the study of viral quasi-species and in the development of the outcomes of genetic engineering where the rate of evolution is much faster compared to natural evolution.

II. Protein Communication Channel

Assuming a first-order Markov channel, the protein communication channel is characterized by the probability transition matrix, $Q = \{q_{i,j}\}_{1 \le i,j \le 20}$, of the amino acids. In this paper, we use two different probability transition matrices: Dayhoff's Point Accepted Mutation (PAM) matrices [7], and a first-order

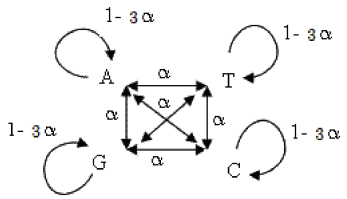


Fig. 4. A probability model of base interchange of any one nucleotide. Here α is the point mutation rate.

Markov probability transition matrix **P** that we build from the genetic code [2]. An element of a PAM matrix, q_{ij}, gives the probability that the amino acid in row i will be replaced by the amino acid in column j after a given evolutionary interval, which is interpreted as the evolutionary distance of the PAM matrix [7]. One PAM would correspond to 1% divergence in a protein, i.e., one amino acid replacement per hundred. The PAM matrices were calculated using mutation data accumulated from phylogenetic trees of closely related sequences. The second probability transition matrix, **P**, is a first-order Markov matrix constructed from the genetic code using a point mutation rate α, which represents the probability of base interchange of any one nucleotide as shown in Fig. 4. The entries of **P** are computed using Baye's rule and assuming that the 64 codons are equally probable. Then, the probability of a transition from amino acid a to amino acid $\hat{a}$ is given by

$$\begin{aligned} \Pr(\hat{a}|a) &= \Pr(\{c_1,\cdots,c_n\}|\{b_1,\cdots,b_m\}) \quad (1)\\ &= \frac{1}{m}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha(k)^{h(b_j,c_i)}(1-3\alpha(k))^{3-h(b_j,c_i)}, \end{aligned}$$

where $\{c_1,\cdots,c_n\}$ (resp. $\{b_1,\cdots,b_m\}$) are the codons of amino acid a (resp. $\hat{a}$), and $h(b_j,c_i)$ is the hamming distance between codon b_j and codon c_i. For computational efficiency and since burst mutations are less likely to happen than one point mutations, we retain only the terms of the first degree in α. For instance, $\Pr(AAC|AAG) = \alpha(1-3\alpha)^2 \simeq \alpha$. The probability transition matrix $\mathbf{P} = \{p_{ij}\}_{1\le i,j\le 20}$ is displayed in Fig. 5. The amino acids are alphabetically ordered by their one-letter standard abbreviations, e.g., $p_{1,1} = \Pr('\text{Alanine}'|'\text{Alanine}') = \Pr(A|A)$.

Since the PAM matrices are constructed using real phylogenetic data, they take into account accepted mutations only, whereas the matrix $\mathbf{P}$ takes into account all possible mutations, whether accepted or rejected by Nature. However, we will show that, for both transition matrices, the biological channel capacity decreases monotonically with time, and therefore genetic information cannot be reliably transmitted after infinitely many generations.

III. Protein Channel Capacity

The capacity of a channel is the maximum rate at which information can be reliably transmitted by the channel. It is defined as [4]

$$C = \max_{p\in P^n} I(p,Q) = \max_{p\in P^n} \sum_j \sum_k p_j Q_{jk} \log \frac{Q_{jk}}{\sum_k p_j Q_{jk}}, \quad (2)$$

where $P^n = \{p \in \mathbb{R}^n : p_j \ge 0\ \forall j; \sum_j p_j = 1\}$ is the set of all probability distributions of the channel input, Q is the probability transition matrix of the channel, and $I(p,Q)$ is known as the mutual information between the channel input and output. The choice of the logarithm base affects the capacity only by a scale factor. In this paper, we use base 2 so that C is expressed in terms of bits-per-channel use.

Evaluation of the channel capacity involves solution of a convex programming problem. In most cases, analytic solutions cannot be found. We therefore adopt the iterative algorithm suggested by Blahut [8] to compute the channel capacity.

Figure 6(a) (resp. 6(b)) shows the channel capacity of the protein communication system channel as a function of the evolutionary distance of PAM matrices (resp. point mutation rate α). As expected, the channel capacity decreases to zero as the evolutionary distance or the point mutation rate α increases. This result has different ramifications on bioinformatics than on communication engineering: In engineering, it is interpreted as a loss of information after a large number of transmissions. The reason is that, in communication engineering, only the initial message is used to convey the information and not the channel. On the other hand, in bioinformatics, the output message captures the information of the channel, i.e. the evolutionary process, regardless of the initial message. In particular, we observe that a parent organism cannot transmit reliably (channel capacity equal to zero) its genetic information to its offspring of many generations no matter how small the point mutation rate, α, is as long as it is non-zero. So, asymptotically, the final distribution of amino acids in offspring depends only on the channel characteristics and not on the parent organism. It is also interesting to observe that organisms with lower mutation rates have higher channel capacity, and therefore their genetic information can be reliably transmitted at a higher rate.

Once we have computed the capacity of the protein communication channel, a comparison of the genomic transmission rate with the channel capacity will reveal whether the genomic code is efficient from an information-theoretic perspective.

IV. Protein Rate Distortion

The rate distortion function, $R(D)$, is the effective rate at which the source produces information subject to the constraint that the receiver can tolerate an average distortion D [5]. A distortion matrix with elements $\rho_{i,j}$ specifies the distortion associated with reproducing the i^{th} source letter by the j^{th} reproducing letter. The rate-distortion function for a discrete memoryless source is defined as [4]

$$R(D) = \min_{Q\in Q_D} I(p,Q) = \min_{Q\in Q_D} \sum_j \sum_k p_j Q_{jk} \log \frac{Q_{jk}}{\sum_k p_j Q_{jk}}, \quad (3)$$

where $Q_D = \{Q \in \mathbb{R}^n \times \mathbb{R}^n : \sum_k Q_{jk} = 1, Q_{jk} \ge 0, d(Q) \le D\}$, $d(Q) = \sum_j \sum_k p_j Q_{jk} \rho_{jk}$, and $p = \{p_j\}$ is the probability vector of the channel input.

A. Evolutionary Process: Mutation and Crossover

It is widely accepted today that the main driving forces of evolution are mutations and unequal crossover[1] [9]. Furthermore, current evidence suggests that Prokaryotes rely mostly on mutations for adaptability and survival, whereas Eukaryotes rely mostly on unequal crossovers [10] [11]. Specifically, it is considered that unequal crossover and other molecular interactions such as gene conversion are contributing to the evolution of multigene families which exist in Eukaryotic genomes [11], and hence govern the evolution of Eukaryotes. On the other hand, Prokaryotes, whose genetic information is encoded in the less stable RNA[2], are more prone to mutations [10]. Consequently, we can fairly postulate that mutations drive the evolution of Prokaryotes whereas unequal crossovers drive the evolution of Eukaryotes. In what follows, we propose an evolutionary model of genomic sequences based on mutations and unequal crossovers, and investigate the average distortions in Prokaryotes versus Eukaryotes.

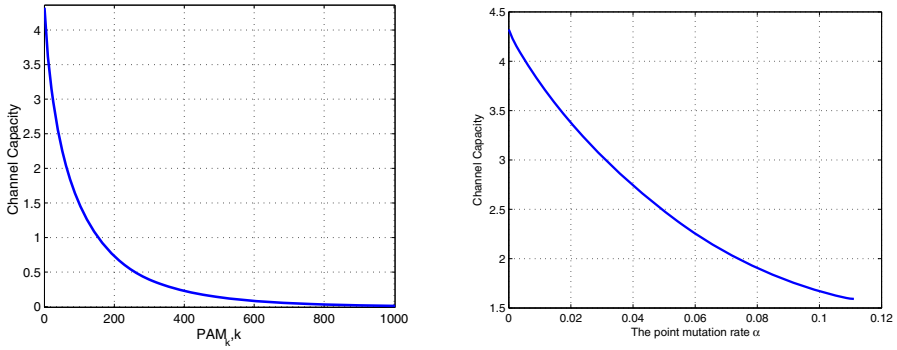
In order to analytically investigate the effects of mutations on the genomes of Prokaryotes and the effects of unequal crossovers on the genomes of Eukaryotes, we need to adopt a probabilistic model that characterizes their occurrences within

[1]Unequal crossover is a crossover between homologous chromosomes that are not perfectly aligned. It results in a duplication of genes on one chromosome and a deletion of these on the other.

[2]Many scientists have pointed out that DNA is more stable than RNA and less prone to mutations [10].

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A	$1-6\alpha$	0	$\alpha/2$	$\alpha/2$	0	α	0	0	0	0	0	0	$\boldsymbol{\alpha}$	0	0	α	α	α	0	0
C	0	$1-7\alpha$	0	0	α	α	0	0	0	0	0	0	0	0	α	2α	0	0	α	α
D	α	0	$1-8\alpha$	2α	0	α	α	0	0	0	0	α	0	0	0	0	0	α	0	α
E	α	**0**	2α	$1-7\alpha$	0	α	0	0	α	0	0	0	0	α	0	0	0	α	0	0
F	0	α	0	0	$1-8\alpha$	0	0	α	0	3α	0	0	0	0	0	α	0	α	0	α
G	α	$\alpha/2$	$\alpha/2$	$\alpha/2$	0	$1-23\alpha/4$	0	0	0	0	0	0	0	0	$3\alpha/2$	$\alpha/2$	0	α	$\alpha/4$	0
H	0	0	α	0	0	0	$1-8\alpha$	0	0	α	0	α	α	2α	α	0	0	0	0	α
I	0	0	0	0	$2\alpha/3$	0	0	$1-7\alpha$	$\alpha/3$	$4\alpha/3$	α	$2\alpha/3$	0	0	$\alpha/3$	$2\alpha/3$	α	α	0	0
K	0	0	0	α	0	0	0	$\alpha/2$	$1-7\alpha$	0	$\alpha/2$	2α	0	α	α	0	α	0	0	0
L	0	0	0	0	α	0	$\alpha/3$	$2\alpha/3$	0	$1-11\alpha/2$	$\alpha/3$	0	$2\alpha/3$	$\alpha/3$	$2\alpha/3$	$\alpha/3$	0	α	$\alpha/6$	0
M	0	0	0	0	0	0	0	3α	α	2α	$1-9\alpha$	0	0	0	α	0	α	α	0	0
N	0	0	α	0	0	**0**	α	α	2α	0	0	$1-8\alpha$	0	0	0	α	α	0	0	α
P	α	0	0	0	0	0	$\alpha/2$	0	0	α	0	0	$1-6\alpha$	$\alpha/2$	α	α	α	0	0	0
Q	0	0	0	α	0	0	2α	0	α	α	0	0	$\boldsymbol{\alpha}$	$1-7\alpha$	α	0	0	0	0	0
R	0	$\alpha/3$	0	0	0	α	$\alpha/3$	$\alpha/6$	$\alpha/3$	$2\alpha/3$	$\alpha/6$	0	$2\alpha/3$	$\alpha/3$	$1-17\alpha/3$	α	$\alpha/3$	0	$\alpha/3$	0
A	$2\alpha/3$	$2\alpha/3$	0	0	$\alpha/3$	$\alpha/3$	0	$\alpha/3$	0	$\alpha/3$	0	$\alpha/3$	$2\alpha/3$	0	α	$1-37\alpha/6$	α	0	$\alpha/6$	$\alpha/3$
R	α	0	0	0	0	0	0	$3\alpha/4$	$\alpha/2$	0	$\alpha/4$	$\alpha/2$	α	0	$\alpha/2$	$3\alpha/2$	$1-6\alpha$	**0**	0	0
V	α	0	$\alpha/2$	$\alpha/2$	$\alpha/2$	α	0	$3\alpha/4$	0	$3\alpha/2$	$\alpha/4$	0	0	0	0	0	0	$1-6\alpha$	0	0
W	0	2α	0	0	0	α	0	0	0	α	0	0	0	0	2α	α	0	0	$1-7\boldsymbol{\alpha}$	0
Y	0	α	α	0	α	0	α	0	0	0	0	α	0	0	0	α	0	0	0	$1-6\alpha$

Fig. 5. **P**: a first-order Markov probability transition matrix between amino acids based on the nucleotide mutation rate α.The amino acids are labeled by their one-letter standard abbreviations. Only the terms of the first degree in the point mutation rate α are retained. An element in **P**, p_{ij}, gives the probability that the amino acid in row i will be replaced by the amino acid in column j according to Eq. (1).



(a) (b)

Fig. 6. Channel Capacity of the protein communication channel: (a) Channel capacity v.s. the evolutionary distance of PAM matrices; (b) Channel capacity v.s. the point mutation rate α. The channel capacity decreases to zero as the evolutionary distance or the point mutation α increases.

the genomes. We propose to model mutations and unequal crossovers as Non-Homogeneous Poisson processes (NHPP) with rate parameters $\lambda_M(t)$ and $\lambda_C(t)$, respectively. A Non-Homogeneous Poisson process (NHPP), $\{N(t) : t \geq 0\}$, is a Poisson process with rate parameter $\lambda(t)$ such that the rate parameter of the process is a function of time [12]. The probability that there are n events in the interval $(r, r+s)$ is given by [12]

$$P\{N(r+s)-N(r)=n\} = \frac{(\int_r^{r+s}\lambda(t))^n e^{-\int_r^{r+s}\lambda(t)\,dt}}{n!} \quad (4)$$

This choice is justified by numerous arguments. First, the Poisson distribution is the limiting distribution of the binomial distribution when the probability of error is small and the genome size is large (De Moivre-Laplace Theorem) [13]. Second, many rare random phenomena in Nature follow a Poisson distribution, e.g., the number of winning tickets in a large lottery, the number of printing errors in a book, etc.

Let us denote by $\{N^{Pr}(t) : t \geq 0\}$ the NHPP with rate parameter $\lambda_M(t)$ modeling the frequency of mutations in Prokaryotic genomes during the interval $(0,t)$, and $\{N^{Eu}(t) : t \geq 0\}$ the NHPP with rate parameter $\lambda_C(t)$ modeling the frequency of unequal crossovers in Eukaryotic genomes during the same interval. Since burst mutations are less likely to happen than one point mutations and for computational efficiency, we take into account only one point mutations. On the other hand unequal crossovers affect an entire segment of the genome.

The distortions of Prokaryotes and Eukaryotes at time t, denoted by $D^{Pr}(t)$ and $D^{Eu}(t)$ respectively, are then given by the sums of the lengths of nucleotide sequences involved in mutations or unequal crossovers during the interval $(0,t)$.

$$D^{Pr}(t) = \sum_{i=0}^{N^{Pr}(t)} u_M(i), \quad (5)$$

$$D^{Eu}(t) = \sum_{i=0}^{N^{Eu}(t)} u_C(i), \quad (6)$$

where $u_M(i) \in \{0,1\}$ (since we are taking into account point mutations only) and $u_C(i) \in \mathbb{N}$, with the assumption that $u_M(0) = 0$ and $u_C(0) = 0$. The average distortions in

Prokaryotes and Eukaryotes at time t are then given by

$$\begin{aligned} E[D^{Pr}(t)] &= E[\sum_{i=0}^{N^{Pr}(t)} u_M(i)] \\ &= \sum_{n=0}^{\infty} n \cdot \frac{e^{-\int_0^t \lambda_M(x)dx}(\int_o^t \lambda_M(x)dx)^n}{n!} \\ &= \int_0^t \lambda_M(x)dx. \end{aligned} \tag{7}$$

$$\begin{aligned} E[D^{Eu}(t)] &= E[\sum_{i=0}^{N^{Eu}(t)} u_C(i)] \\ &= \sum_{k=1}^{\infty}\sum_{n=0}^{\infty} n \cdot l_k \cdot \frac{e^{-\int_0^t p_k\lambda_C(x)dx}(\int_0^t p_k\lambda_C(x)dx)^n}{n!} \end{aligned} \tag{8}$$

$$\begin{aligned} &= \sum_{k=1}^{\infty} l_k \cdot \int_0^t p_k\lambda_C(x)dx \\ &= \sum_{k=1}^{\infty} l_k p_k \cdot \int_0^t \lambda_C(x)dx \\ &= \eta_l \cdot \int_0^t \lambda_C(x)dx, \end{aligned} \tag{9}$$

where Eq. (8) follows from the Representation Theorem for a Compound Poisson Process having a denumerable mark space [Chapter 4] [12], l_k is the length of the genome segment affected by an unequal crossover with probability p_k, and

$$\eta_l = \sum_{k=1}^{\infty} l_k p_k, \tag{10}$$

is the average length of nucleotide sequences affected by unequal crossovers. Observe that Eq. (7) is a special case of Eq. (9) corresponding to $\eta_l = 1$.

In what follows, we consider a multi-exponential model for the rate parameter $\lambda(t)$:

$$\lambda(t) = \sum_{k=1}^{K} \lambda_k(t), \tag{11}$$

with the rate parameter of the k^{th} model compartment being

$$\lambda_k(t) = a_k e^{-b_k t}, \quad t \geq 0, \tag{12}$$

where $\{a_k, b_k : k = 1, \cdots, K\}$ are real parameters. Due to the multi-exponential nature of the long time behavior of various biological phenomena, multi-exponential models have been widely used in numerous areas of medicine and biology. For instance, the multi-exponential model has been used to describe the time-dependent behavior of a radioactive tracer in physiochemical studies in nuclear medicine [14]. It has also been used in describing long time dynamics in proteins [15], such as the kinetics of protein folding [16], and the tryptophan fluorescence intensity decay [17]. We propose to use a multi-exponential model that consists of two model compartments, that is

$$\lambda_M(t) = a_m - a_m e^{-k_m t}, \tag{13}$$

$$\lambda_C(t) = b_c - b_c e^{-k_c t}, \tag{14}$$

where a_m, k_m, b_c and k_c are positive parameters. Observe that this model reflects a speed of evolution that increases with time till it reaches an asymptotic upper bound. Here the upper bounds of evolutionary speed are a_m and b_c for mutations and unequal crossovers, respectively. k_m and k_c represent the evolutionary acceleration. We assume that $k_m < k_c$ since unequal crossovers involve longer nucleotide sequences and hence lead to higher distortions compared to mutations. By substituting Eqs. (13) and (14) into Eqs. (7) and (9), we obtain

$$\begin{aligned} E[D^{Pr}(t)] &= \int_0^t (a_m - a_m e^{-k_m x})dx \\ &= a_m t + \frac{a_m}{k_m} e^{-k_m t} - \frac{a_m}{k_m}. \end{aligned} \tag{15}$$

$$\begin{aligned} E[D^{Eu}(t)] &= \eta_l \cdot \int_0^t (b_c - b_c e^{-k_c x})dx \\ &= \eta_l b_c t + \eta_l \frac{b_c}{k_c} e^{-k_c t} - \eta_l \frac{b_c}{k_c}. \end{aligned} \tag{16}$$

It can be easily shown that

$$E[D^{Pr}(t)] \geq E[D^{Eu}(t)], \quad \text{if } a_m \geq \eta_l b_c k_c / k_m, \tag{17}$$

$$E[D^{Pr}(t)] \leq E[D^{Eu}(t)], \quad \text{if } a_m \leq \eta_l b_c, \tag{18}$$

for all $t \geq 0$. If $\eta_l b_c < a_m < \eta_l b_c k_c / k_m$, the curves intersect. In general, a comparison between the two curves requires a search in the five-dimensional parameter space $S = \{a_m, k_m, b_c, k_c, \eta_l\}$. However, we can study the effect of one parameter on the average distortion curves of Eukaryotes and Prokaryotes by varying that parameter while fixing the other four. For instance, we can investigate the effect of mutation accumulation in Prokaryotic genomes by varying the parameter a_m while fixing k_m, b_c, k_c and η_l. To this purpose, we consider three parameter sets, S_0, S_1 and S_2, that differ only in the value of a_m. We set $S_0 = \{2, 0.01, 0.01, 0.04, 300\}$, $S_1 = \{4, 0.01, 0.01, 0.04, 300\}$ and $S_2 = \{12, 0.01, 0.01, 0.04, 300\}$. Figure 7 shows the rate parameter curves of Eukaryotes, $\lambda_C(t)$, and Prokaryotes, $\lambda_M(t)$, corresponding to these parameter sets (see Eqs. (13) and (14)). Since $\lambda_C(t)$ is independent of a_m, the rate parameter curves of Eukaryotes corresponding to S_0, S_1 and S_2 are identical. The rate parameter curve of Prokatyotes, however, is proportional to a_m, and hence increases for larger values of this parameter. Observe that the rate parameter of Eukaryotes, $\lambda_C(t)$, is much smaller than the one of Prokaryotes, $\lambda_M(t)$, i.e., $\lambda_C(t) << \lambda_M(t)$, for all $t \geq 0$. This is true in reality since unequal-crossover is a rare phenomenon, whereas mutations occur at almost every cell cycle. Figure 8 shows the average distortion curves of Eukaryotes, $E[D^{Eu}(t)]$, and Prokaryotes, $E[D^{Pr}(t)]$, corresponding to S_0, S_1 and S_2 (see Eqs. (15) and (16)). Since $E[D^{Eu}(t)]$ is independent of a_m, the rate-distortion curves of Eukaryotes corresponding to these three parameter sets are identical. The average distortion curve of Prokaryotes, however, increases with a_m as follows: for small values of a_m, it is below the curve of Eukaryotes, so that Prokaryotes always have less distortion than Eukaryotes; whereas for large values of a_m, the distortion in

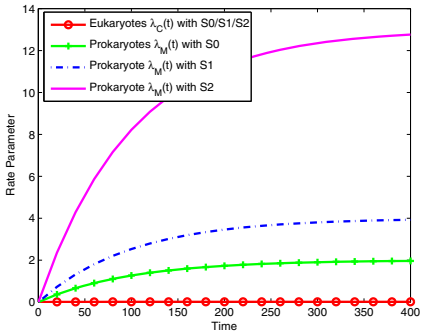


Fig. 7. The rate parameter curves of Eukaryotes, $\lambda_C(t)$, and Prokaryotes, $\lambda_M(t)$, corresponding to the three parameter sets $S_0 = \{2, 0.01, 0.01, 0.04, 300\}$, $S_1 = \{4, 0.01, 0.01, 0.04, 300\}$ and $S_2 = \{12, 0.01, 0.01, 0.04, 300\}$. Observe that since only the parameter a_m is varied between S_0, S_1 and S_2, the rate parameter curves of Eukaryotes, $\lambda_C(t)$, corresponding to S_0, S_1 and S_2 are identical.

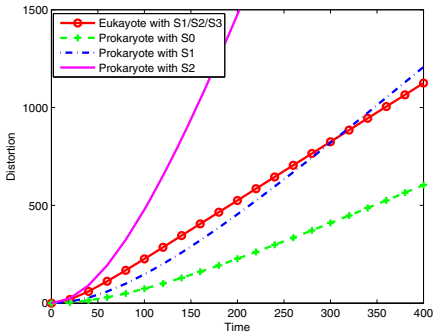


Fig. 8. Time-distortion curves of Eukaryotes and Prokaryotes for the parameter sets $S_0 = \{2, 0.01, 0.01, 0.04, 300\}$, $S_1 = \{4, 0.01, 0.01, 0.04, 300\}$ and $S_2 = \{12, 0.01, 0.01, 0.04, 300\}$. Here, the curves of Eukaryotes corresponding to these three parameter sets are identical.

Prokaryotes becomes higher than Eukaryotes. When $\eta_l b_c < a_m < \eta_l b_c k_c / k_m$, the two curves intersect. The distortion can be associated with the evolutionary distance. That is, low distortion regions would correspond to small evolutionary distances, whereas high distortion regions would correspond to larger evolutionary distances. It is then quite interesting to observe that, for small evolutionary distances (in particular at the beginning of life), Prokaryotes are more efficient than Eukaryotes from an average distortion point of view, and for larger evolutionary distances, Eukaryotes become more efficient. Moreover, at some point in time, the two curves intersect and hence the average distortions of Prokaryotes and Eukaryotes are equal at that time. The evolution of the average distortion curves of Prokayotes and Eukaryotes can intuitively be explained as follows: since a mutation affects a considerably smaller number of nucleotides than an unequal crossover, it induces less modification to the genome sequence. Thus, at the beginning of evolution, we expect the distortion induced by mutations to be smaller compared to the distortion induced by unequal crossovers. However, with time, mutations accumulate much faster than unequal crossovers, which happen very infrequently ($\lambda_M(t) >> \lambda_C(t)$). Consequently, the distortion induced by mutations exceeds, over time, the distortion induced by unequal crossover. This implies higher fidelity, over time, in Eukaryotes than Prokaryotes.

Although different values of the parameter set S may result in different relationships between the average distortion curves of Prokaryotes and Eukaryotes, we are assured that Nature has already chosen the "adequate" parameter set for its evolutionary process.

B. Rate-Distortion Curves of the Three Domains of Life

In this section, we use Blahut's algorithm for rate-distortion functions to experimentally compute the rate distortion curves (see Eq. (3)) of the three domains of life: Archaea, Bacteria and Eukaryotes. The amino acid probability distributions in Archaea, Bacteria and Eukaryotes were computed in [18] based on real data. We define the distortion between a pair of amino acids as their L_2 distance in the 2-D Principal Component Analysis (PCA) plane, shown in Fig 10, where the amino acids are labeled by their one-letter standard abbreviations. The PCA plane is obtained from the 7-D space which is characterized by the following 7 physico-chemical properties: volume, bulkiness, polarity, PH index, hydrophobicity scale, surface and fractional area. These properties are important in determining protein structure and are obtained from [Chapter 2] [19]. The further the distance between any two amino acids on the PCA plane, the more alteration introduced to protein structure when substitution between these two amino acids happens (due to mutations or unequal crossovers).

Figure 9 shows the rate-distortion curves for Archaea, Bacteria and Eukaryotes. It reveals two distinct regions: a low distortion region ($0 \leq D \leq 1.4$) and a high distortion

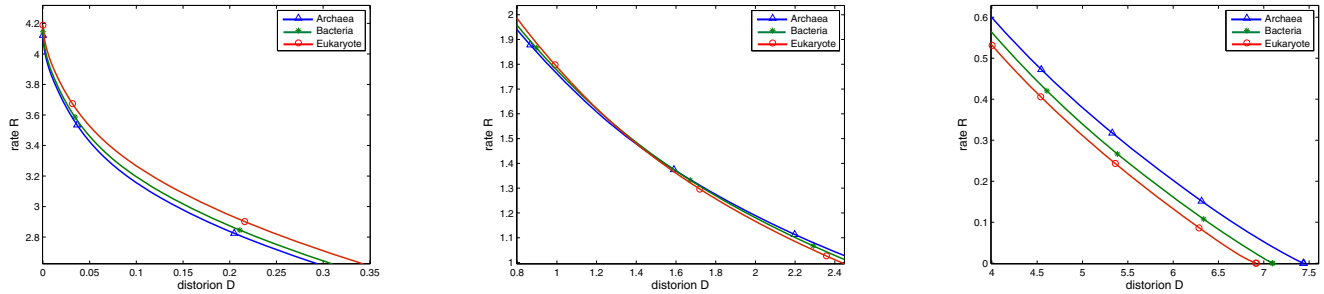


Fig. 9. Rate-distortion curves for Archaea, Bacteria and Eukaryotes: (a) low distortion region; (b) intersection region; (c) high distortion region. In the low distortion region, the R-D curve of Eukaryotes is the highest followed by Bacteria, then Archaea. The three R-D curves intersect at the point $(D_0, R_0) \approx (1.4, 1.5)$, then reverse their order.

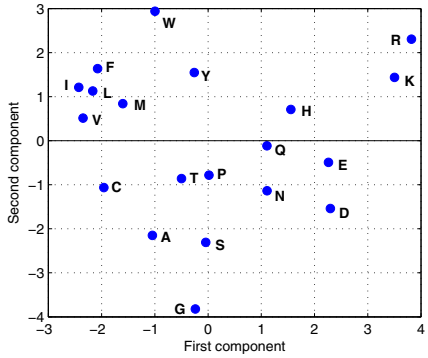


Fig. 10. Plot of the amino acids on the first two components of the PCA space. The amino acids are labeled by their one-letter standard abbreviations. The distortion between a pair of amino acids is defined as their L_2 distance on the PCA plane.

region ($1.4 \leq D \leq 7.5$). In the low-distortion region, the R-D curve of Eukaryotes is the highest followed by Bacteria, then Archaea, i.e.,

$$R(D)_{Ar} < R(D)_{Ba} < R(D)_{Eu}, \quad \forall\ 0 < D < 1.4, \tag{19}$$

where $R(D)_{Ar}, R(D)_{Ba}$ and $R(D)_{Eu}$ denote the rate-distortion curves of Archaea, Bacteria and Eukaryotes, respectively. At about $D \approx 1.4$, the above order switches to

$$R(D)_{Eu} < R(D)_{Ba} < R(D)_{Ar}, \quad \forall\ 1.4 < D < 7.5. \tag{20}$$

We observe that for small distortions or small evolutionary distances, Archaea was the most efficient organism from an information-theoretic perspective, followed by Bacteria and then Eukaryotes. Specifically, given a fixed transmission rate (of the genetic information), Archaea would have the least distortion whereas Eukaryotes would have the greatest. At the point $(D_0, R_0) \approx (1.4, 1.5)$, the three R-D curves intersect and reverse their order. Thus, for large evolutionary distances, Eukaryotes maintain the greatest biological fidelity among the three domains of life. This experimental result validates the evolutionary model proposed in the previous section.

The actual average distortion over the protein communication channel is defined as

$$D = \sum_j \sum_k p_j q_{jk} \rho_{jk}, \tag{21}$$

where $Q = \{q_{i,j}\}$ is the probability transition matrix of the channel, $p = \{p_j\}$ is the probability vector of the channel input and $\rho_{i,j}$ is the distortion between amino acids i and j. We use PAM_{250} as the probability transition matrix of the channel. Dayhoff et al. [7] found that the PAM_{250} matrix works well for scoring actual protein sequences. This evolutionary distance corresponds to 250 substitutions per hundred residues in a protein sequence. The actual average distortions for Archaea, Bacteria and Eukaryotes are displayed in Table I. Observe that the biological rate-distortion values $R(D)$, corresponding to the average distortions given in Table I, are less than the Shannon channel capacity ($C = 0.8197 > R(D)$). Therefore, we can ascertain, from a rate-distortion theory viewpoint, that the genetic information is encoded such that the system reproduces the initial input with fidelity D. In particular, the biological communication system does not rely on codes that produce negligible errors since the level of distortion presented must account for evolutionary processes.

Finally, it is important to observe that the formula of rate-distortion function given in Eq. (3) is valid only for discrete-time memoryless sources. For discrete-time stationary sources with memory, Wyner and Ziv [20] derived bounds for the rate-distortion function, $R^*(D)$, as follows:

$$R(D) - \Delta \leq R^*(D) \leq R(D)\ , \tag{22}$$

where $R(D)$ is the rate-distortion function of the memoryless source with the same marginal statistics, Δ is a measure

TABLE I
AVERAGE RATE-DISTORTION FOR THREE DOMAINS OF LIFE

	Archaea	Bacteria	Eukaryote
Distortion	9.1491	8.9964	8.8979

of the memory of the source and is independent of the distortion measure and the distortion value D. Thus, the R-D curves for discrete-time stationary sources with memory are always shifted down compared to the R-D curves of the corresponding memoryless sources. Moreover, the shift is only a function of the source and not the distortion. Thus, the biological rate-distortion values $R^*(D)$ corresponding to the average distortions given in Table I are still less than the Shannon channel capacity and the evolution of the rate-distortion curves of Archaea, Bacteria and Eukaryotes would still exhibit the same reversal phenomenon depicted in Fig. 9.

V. CONCLUSION

By modeling the evolutionary process as the iteration of a protein communication system over time, we were able to study it from an information-theoretic perspective. Investigation of the biological communication channel capacity and rate-distortion curves of the three domains of life: Archaea, Bacteria and Eukaryotes, reveals that the biological fidelity D does not reach the Shannon optimum distortion. Furthermore, we relied on these results to provide an evolutionary model of these three branches of life based on mutations and unequal crossovers. The proposed evolutionary model has been shown to provide a possible explanation of the pattern of rate-distortion bounds for the basic life forms. Moreover, the model is consistent with existing evolutionary dynamic theory: high initial mutation and crossover rates that moderated over time [21]. This phenomenon could be explained, in the case of Eukaryotes, based on two distinct principles: (a) the stability of an organism increases as the length of the introns increases since the frequency of gradual changes due to mutations and crossover in the exons decreases [2]; and (b) adaptability by evolutionary jumps due to unequal crossover rises as the length of the introns increases according to the gene-shuffling theory introduced by Gilbert [22] [23] [24]. Thus, a gradual decrease in mutation and crossover rate [21] coupled with evidence of larger evolutionary jumps [25] [26] [27] (or punctuated equilibria [28] [29] [30]) is consistent with the appearance and gradual increase in the length of introns in the genomic sequence. Finally, the proposed theoretical model could provide a mathematical framework for the study of viral quasi-species and the development of the outcomes of genetic engineering where the rate of evolution is much faster compared to natural evolution.

REFERENCES

[1] A. Bruce, D. Bray, J. Lewis, M. Raff, K. Roberts, and J. D. Watson, *Molecular Biology of the Cell*, Garland Publishing, New York, 3rd edition, 1994.

[2] N. Bouaynaya and D. Schonfeld, "Protein communication system: Evolution and genomic structure," *Algorithmica*, vol. 48, no. 4, pp. 375–397, August 2007.

[3] C. Shannon, "A mathematical theory of communication," *Bell System Technical Journal*, vol. 27, pp. 379–423, July 1948.

[4] Thomas M Cover and Joy A Thomas, *Elements of information theory*, Wiley-Interscience, 1991.

[5] Toby Berger, *Rate distortion theory: A mathematical basis for data compression*, Prentice Hall, New Jerzy, 1971.

[6] L. Gong, N. Bouaynaya, and D. Schonfeld, "Information-theoretic bounds of evolutionary processes modeled as a protein communication system," in *IEEE Statistical Signal Processing Workshop*, Madison, WI, August 2007, pp. 1–5.

[7] M. Dayhoff, R. Schwartz, and B. Orcutt, "A model of evolutionary change in proteins," *Altas of Protein Sequence and Structure*, vol. 5, no. 3, pp. 345–352, 1978.

[8] R. Blahut, "Computation of channel capacity and rate-distortion functions," *IEEE Transactions on Information Theory*, vol. 18, no. 4, pp. 460–472, July 1972.

[9] T. Ohta, "Evolution of gene families," *Gene*, vol. 259, pp. 45–52, 2000.

[10] C. Zimmer, "Did DNA come from viruses," *Science*, vol. 312, no. 5775, pp. 870–872, May 2006.

[11] T. Ohta, "Simulating evolution by gene duplication," *Genetics*, vol. 115, no. 1, pp. 207–213, Jan 1987.

[12] D. Snyder and M. Miller, *Random Point Processes in Time and Space*, Springer, New York, 2nd edition, 1991.

[13] A. Papoulis and S. U. Pillai, *Probabilities, Random Variables and Stochastic Processes*, McGraw-Hill, 1991.

[14] J. A. Jacques, "Tracer kinetics," in *Principles of Nuclear Medicine*. H. N. Wagner, etc, Philadelphia, 1968.

[15] X. Song X. J. Jordanides, M. J. Lang and G. R. Fleming, "Solvation dynamics in protein environments studied by photon echo spectroscopy," *Journal of Physical Chemistry B*, vol. 103, no. 37, pp. 7995–8005, 1999.

[16] J. Ervin J. Sabelko and M. Gruebele, "Observation of strange kinetics in protein folding," *Proceedings of the National Academy of Sciences USA*, vol. 96, no. 11, 1999.

[17] Z. Bajzer and F. G. Prendergast, "A model for multiexponential tryptophan fluorescence intensity decay in proteins," *Biophysical Journal*, vol. 65, no. 6, pp. 2313–2323, 1993.

[18] N. Bogatryreva, A. Finkelstein, and O. Galzitskaya, "Trend of amino acid composition of proteins of different taxa," *Journal of Bioinformatics and Computational Biology*, vol. 4, no. 2, pp. 597–608, 2006.

[19] P. Higgs and T. Attwood, *Bioinformatics and Molecular Evolution*, Blackwell publishing, 2005.

[20] A. Wyner and J. Ziv, "Bounds on the rate-distortion function for stationary sources with memory," *IEEE Transactions on Information Theory*, vol. 17, no. 5, pp. 508–513, September 1971.

[21] C. Woese, "The universal ancestor," *Proceedings of the National Academy of Sciences*, vol. 95, no. 12, pp. 6854–6859, June 1998.

[22] W. Gilbert, "Why genes in pieces?," *Nature*, vol. 271, pp. 501, February 1978.

[23] W. Gilbert, "Gene structure and evolutionary theory," in *New Perspective on Evolution*, 1991, pp. 155–163.

[24] W. Gilbert, S. D. Souza, and M. Long, "Origin of genes," *Proceedings of the National Academy of Sciences*, vol. 94, no. 15, pp. 7698–7703, July 1997.

[25] S. J. Gould, "Evolution's erratic pace," *Natural History*, vol. 86, no. 5, pp. 14, May 1977.

[26] N. W. Gillham, "Evolution by jumps: Francis galton and william bateson and the mechanism of evolutionary change," *Genetics*, vol. 159, no. 4, pp. 1383–1392, December 2001.

[27] J. Zhang, D. M. Webb, and O. Podlaha, "Accelerated protein evolution and origins of human-specific features: Foxp2 as an example," *Genetics*, vol. 162, no. 4, pp. 1825–1835, December 2002.

[28] N. Eldredge and S. J. Gould, "Punctuated equilibria: an alternative to phyletic gradualism," in *Models in Paleobiology*, T. J. M. Schopf, Ed., San Francisco, 1972, Freeman, pp. 82–115.

[29] E. Mayr, "Speciational evolution or punctuated equilibria," in *The Dynamics of Evolution*, A. Somit and S. A. Peterson, Eds., New York, 1992, pp. 21–48, Cornel University Press.

[30] S. J. Gould, *The Structure of Evolutionary Theory*, Harvard University Press, Cambridge, MA, 2002.

Liuling Gong was born in Xinyu, China, in 1986. She received the B.S. degree in Electronic Engineering and Information Science from the University of Science and Technology of China (USTC), Hefei, China, in 2006. She is currently pursuing the Ph.D. degree at the Department of Electrical and Computer Engineering, University of Illinois at Chicago.

Her research interests include signal processing, machine learning, data mining, and statistical modeling.

Nidhal Bouaynaya received the B.S. degree in Electrical and Computer Engineering from the Ecole Nationale Suprieure de l'Electronique et de ses Applications (ENSEA), France; and the M.S. degree in Electrical and Computer Engineering from the Illinois Institute of Technology, Chicago, IL, in 2002; the Dipl?me d'Etudes Approfondies in signal and image processing from ENSEA, France, in 2003; the M.S. degree in Mathematics and the Ph.D. degree in Electrical and Computer Engineering from the University of Illinois at Chicago, Chicago, IL, in 2007.

In fall 2007, she joined the University of Arkansas at Little Rock where she is currently an Assistant Professor in the Department of Systems Engineering. Dr. Bouaynaya won the Best Student Paper Award in Visual Communication and Image Processing 2006.

Her research interests are in signal, image and video processing; mathematical morphology and genomic signal processing.

Dan Schonfeld received the B.S. degree in Electrical Engineering and Computer Science from the University of California, Berkeley, California, and the M.S. and Ph.D. degrees in Electrical and Computer Engineering from the Johns Hopkins University, Baltimore, Maryland, in 1986, 1988, and 1990, respectively.

In August 1990, he joined the Department of Electrical Engineering and Computer Science at the University of Illinois, Chicago, Illinois, where he is currently a Professor in the Departments of Electrical and Computer Engineering, Computer Science, and Bioengineering, and Co-Director of the Multimedia Communications Laboratory (MCL) and member of the Signal and Image Research Laboratory (SIRL).

Dr. Schonfeld has authored over 120 technical papers in various journals and conferences. He was co-author (with Junlan Yang) of a paper that won the Best Student Paper Award in the IEEE International Conference on Image Processing 2007. He was also co-author (with Wei Qu) of a paper that won the Best Student Paper Award in the IEEE International Conference on Image Processing 2006. He was also co-author (with Nidhal Bouaynaya) of a paper that won the Best Student Paper Award in Visual Communication and Image Processing 2006.

Dr. Schonfeld is currently serving as Region 1-6 representative in the Chapters Committee of the IEEE Signal Processing Society. He is also serving as Chairman of the Chicago IEEE Signal Processing Chapter. He is currently serving on the IEEE Image and Multidimensional Signal Processing Technical Committee. He has been elected as a Senior Member of the IEEE.

Dr. Schonfeld is currently serving as special sections area editor for the IEEE Signal Processing Magazine. He is currently also serving as associate editor of the IEEE Transactions on Image Processing on Image and Video Storage, Retrieval and Analysis and associate editor of the IEEE Transactions on Circuits and Systems for Video Technology on Video Analysis. He has served as an associate editor of the IEEE Transactions on Signal Processing on Multidimensional Signal Processing and Multimedia Signal Processing as well as an associate editor of the IEEE Transactions on Image Processing on Nonlinear Filtering.

Dr. Schonfeld has served as chair of the SPIE Conference on Visual Communication and Image Processing 2007. He was a member of the organizing committees of the IEEE International Conference on Image Processing 1998 and 2012 and IEEE Workshop on Nonlinear Signal and Image Processing 1997. He was also the plenary speaker at the INPT/ASME International Conference on Communications, Signals, and Systems in 1995 and 2001.

His current research interests are in multi-dimensional signal processing; image and video analysis; computer vision; and genomic signal processing.