Algorithmica (2007) 48: 375–397
DOI: 10.1007/s00453-007-0180-6

Algorithmica

Protein Communication System: Evolution and Genomic Structure

Nidhal Bouaynaya[1] and Dan Schonfeld[1]

Abstract. We develop a mathematical model of the genetic information storage and transmission system and investigate its properties. Breaking with tradition, whereas the genetic information storage and transmission apparatus is conventionally modelled as an engineering communication system with the DNA sequence as the input and the amino acid chain as the output, in this paper the genetic communication model is viewed as one between proteins. A connection in a series of protein communication systems is equivalent to a channel through time: the Channel of Evolution. We investigate the dynamics of the channel of evolution in both cases of a constant and time-variant point mutation rates. We prove that the distribution of amino acids converges geometrically to a specific distribution which matches nearly perfectly an estimate of the natural abundance of amino acids in Nature today. Moreover, based on the highly redundant structure of the encoded genetic message (i.e., DNA), we demonstrate that the role of introns in eukaryotic genomes is to maintain a fine balance between two competing yet complementary forces: stability and adaptability. The stability role is evaluated by showing that introns play the role of a decoy in absorbing mutations. We derive the optimal exon length distribution, which minimizes the probability of error in eukaryotic genomes. Furthermore, to understand how Nature can physically achieve such a distribution, we propose a diffusive random walk model for exon generation throughout evolution. This model results in an alpha stable distribution, which is asymptotically equivalent to the optimal distribution. Experimental results on various eukaryotic organisms spanning the phylogenetic tree from unicellular organisms to plants to vertebrates show that both distributions accurately fit the biological data.

Key Words. Protein communication, Protein evolution, Amino acid distribution, Genomic structure, Iotron models, Exon models.

1. Introduction.

Over the past half a century we have undergone a revolution in our ability to archive, process and exchange information. Communication of biological systems took a head start 3.5 billion years ago. However, for all the strengthened efforts that are directed towards the study of complex engineered information processing systems, remarkably little is known about the broad role of information in biological systems.

Communication systems are used to study both transmission of information between remote locations and data storage for future retrieval [1]. Information theoretic principals have been used to develop effective algorithms to transmit information successfully from a source to a receiver in engineered systems [2]. Living systems also successfully transmit their genetic information through complex biological processes such as replication, transcription and translation. Moreover, the genetic information storage and transmission system is common to the three domains of life (archae, prokaryotes and eukarytes)

[1] Department of Electrical and Computer Engineering, University of Illinois at Chicago, Chicago, IL 60607-7053, USA. {nbouay1,dans}@uic.edu.

Received March 16, 2006; revised February 26, 2007. Communicated by P. Berman, B. DasGupta, and J. Liang. Online publication July 13, 2007.

just like engineered communication systems are designed for all possible messages regardless of their semantic meanings. The genetic information storage and transmission apparatus resembles communication engineering systems in many ways: the genetic information is encoded in the DNA. By decoding genes into proteins, organisms come into being. However, unlike communication engineer's systems, the genetic communication system is not designed to minimize transmission errors. In the absence of errors, evolution will not be possible. Furthermore, perfect (i.e., errorless) communication of the genetic information spells stagnation and ultimately extinction. So, intuitively, there has to be a balance between maintaining the organism identity by reliable transmission of its genetic information (stability) and allowing errors to occur purposefully to encourage evolution (adaptability). Then what is the right mathematical model to capture the genetic information storage and transmission system? Moreover, can we mathematically quantify Nature's design specifications which balance stability and adaptability? This paper seeks to address these two questions.

Several researchers have explored the central dogma of genetics from an information transmission viewpoint [3]–[8]. Gatlin [3], Yockey [4], [7] and Roldan et al. [5] model the genetic information transfer as a communication system, where the input is the DNA sequence and the output is the amino acid chain in the protein. That is the channel of the genetic communication system is the translation process. May [6], [9] and Rosen [8] consider the channel to be the replication and transcription processes whereas the translation process models the decoder of the system. However, both models are inconsistent with engineering communication systems, which model transmission and storage of the same messages at the source and destination (excluding errors due to channel degradation). As Shannon clearly states in his seminal paper: "The fundamental problem of communication is that of reproducing at one point either exactly or approximately a message selected at another point" [2]. Consequently, the reductionist approach to the central dogma as often misused by "DNA $\rightarrow$ RNA $\rightarrow$ protein" cannot be modelled by a communication engineering system but rather by some non-bijective mathematical transformation or decoder, which maps the 4-letter alphabet message in the DNA into the 20-letter alphabet message in the amino acid polypeptide. Moreover, Yockey's and Rodan's DNA–protein system view the DNA as the source and hence completely miss the true nature of the DNA sequence as the encoded genetic information. In particular, this view accounts for the existence of non-coding segments in the DNA.

In this paper we model the transmission of genetic information as a communication system between proteins. The proposed protein communication system is merely an abstraction, which models a cell as a set of proteins and the process of cell division as an information communication system between protein sets. This model does not support either the theories of proteins-first or nucleotides-first at the origin of life. In fact, the proposed communication model could be used to explain the transmission of information in both the proteins-first and nucleotides-first theories. The encoding process, in the proposed protein communication channel, does not happen in biology since proteins cannot be used to generate DNA. It is only a mathematical model of the protein information captured by DNA. To clarify this idea, assume that we have a computer that maintains an MPEG code while decoding to display a video. Copies of the video to other computers only require sending the MPEG code. Assume further that the first MPEG code was created by chance. This system never encodes a video into

MPEG. It only decodes MPEG to display a video. The proper communication model is, however, "video → MPEG → MPEG → video" even though the process "video → MPEG" never takes place. Biological organisms have resolved the real communication problem, i.e. "protein → protein", by ensuring that organisms maintain both proteins and DNA. Therefore, the "protein → DNA" encoder is not required biologically. Biological systems only decode DNA into proteins via the transcription and translation processes. Furthermore, based on the highly redundant structure of the DNA sequence, i.e., presence of a large percentage of non-coding segments, we argue that the encoder models a source and channel encoder [10].

Even though the genetic encoder is unknown, studying the structure of the DNA sequence might shed light onto the evolutionary constraints which shaped it. In other words, the encoded information (i.e., DNA) should reflect in its structure the biological system design specifications: stability and adaptability. An amazing feature of the DNA is its phenomenal redundancy. The genes of eukaryotic genomes contain protein-coding sequences, called *exons*, seperated by non-coding sequences, called *introns*. Thus, introns are excluded from the main gene function: making proteins. The great deal of extra energy required to sustain, process and conserve introns during many millions of years of evolution may imply an essential function. Otherwise, most likely they would have been eliminated by natural selection long ago. It appears difficult to prove this via molecular biology. A better strategy would be to seek an answer outside of the traditional domain. From a communication engineering point of view, the so-called "junk DNA" may turn out to be just as important as the much sought-after genes. Forsdyke [11] and later Battail [10], [12] hypothesized that error-correcting codes are used in the replication process of the genome. A consequence of this hypothesis is the existence of redundant DNA. The genes in the DNA are viewed as the encoded messages composed of the information symbols (i.e., exons) and the redundant symbols (i.e., introns) needed by the error-correction process. It is well known that DNA replication and protein synthesis involve error repair mechanisms [13]. However, no linkage has been found between these repair mechanisms and the intron sequences in the genes. Liebovitch et al. [14] developed a procedure to check for the existence of a linear block code in genetic sequences. If a linear block error-correcting code is present in DNA then some bases would be a linear function of the other bases in each set of bases. However, their experimental results on the lac operon and the gene for cytochrome c revealed that these two genes do not appear to contain such a simple error-correcting code. So, either there is no error-correction mechanisms encoded in the introns or the genetic error-correcting mechanisms are algorithmically different from what has been tested in the literature so far [15], [12].

We propose that introns control the balance between stability and adaptability in eukaryotic genomes. On the one hand, introns drive evolution by increasing the rate of recombination of exons via unequal crossover (adaptability) [16]. On the other hand, they play the role of a decoy for mutations (stability). For example, recent experiments removed 1% of the mouse genome and were unable to detect any effect on the phenotype [17]. So, the role of introns in increasing the rate of unequal crossovers is tempered in order to prevent excessive evolutionary adaptability. Rapid changes in the the genome must not occur too frequently, or else we would experience evolutionary jumps in each generation. In this paper we prove the stability role of introns based on probability of error analysis and optimization. The stability role attributed to introns accounts for at

least two biological facts:

(i) The absence of introns in prokaryotic genomes translates, according to our view, to a high mutability rate of these primitive organisms. It is widely known today that many bacteria and viruses rely on mutations for diversification.
(ii) The decoy role for introns predicts that coding sequences should be more conserved among organisms than non-coding sequences. Studies in comparative genomics showed that functional DNA sequences tend to undergo mutation at a slower rate than non-functional sequences. For example, the coding sequence of a human protein-coding gene is typically about 80% identical to its mouse ortholog, while their genomes as a whole are much more widely divergent.

This paper is organized as follows: In Section 2 we define the protein communication channel and introduce its probability transition matrix. A series connection of the protein communication channel is equivalent to a channel through time: the channel of evolution. We study the dynamics of the channel of evolution in both cases of constant and time-varying point mutation rates. In Section 3 we formulate an optimization problem to determine the optimal exon length distribution, which minimizes the probability of error in eukaryotic genomes. First, we derive the optimal exon length distribution. Second, we address the question of a feasible physical realization of such a distribution. We show that a diffusive random walk model for exon generation throughout evolution leads to an exon length density, which is asymptotically equivalent to the optimal distribution. In Section 4 we compute the exon length distribution of various eukaryotic organisms spanning the phylogenetic tree from unicellular organisms to invertebrates to vertebrates. Amazingly, the alpha-stable and the optimal distributions accurately fit the empirical exon length distribution of the different eukaryotic organisms. Finally, Section 5 summarizes the main results of this paper and discusses future work.

2. Protein Communication Channel. We model the transmission of information, during cell division or asexual reproduction, as a protein communication system with a single source generating the protein set of the mother cell. The protein communication system is shown in Figure 1. The physical channel models the transmission and storage medium and is the source of errors. Chemical mutagens and radiation cause errors in DNA during storage and replication [18]. The decoder modelled by the transcription and

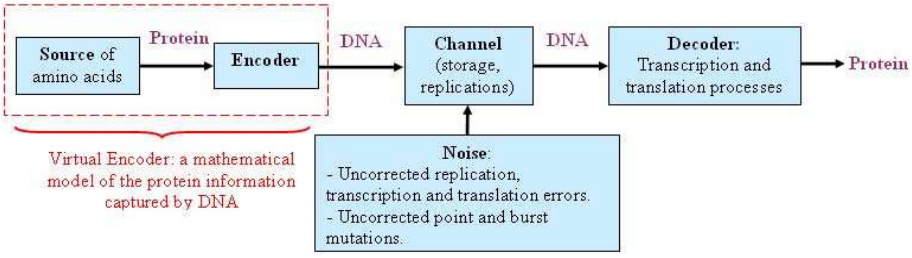

Fig. 1. Protein communication system.

Engineering Communication System	Protein Communication System
Video	Set of proteins of the cell
MPEG	DNA
Encoder	—
Decoder	Translation Process
Objective function: minimize the probability of error	*Objective function*: balance between maintaining the cell's identity by reliable transmission of its protein set and allowing errors to occur purposefully to encourage evolution.

Fig. 2. Comparison between the engineering communication system for video transmission and the protein communication system during cell replication.

translation processes is not free of errors either [19]. However, to simplify the model, these errors are incorporated as part of the channel. Figure 2 summarizes the analogy between an engineering communication system for video transmission and the protein communication system.

A protein communication system, which models the transmission of information in sexual reproduction, is much more involved mathematically than the single source communication system in cell replication. From an information theoretic perspective we have two sources; each source is a parent containing two homologous protein sets. The output of this communication model consists of two proteins, randomly selected from each parent, received after transmission over the communication channel. Analysis of this communication system requires the use of multi-user information theory and distributed coding. For analytical simplicity, we decompose this complex system into two parallel communication systems. Each communication system consists of a source (a single parent) generating two homologous protein sets. A stochastic process selects one protein from each homologous pair. The selected protein is transmitted through an identical communication system to the single source protein channel depicted in Figure 1. The received message is formed by the union of the two proteins received from each parent.

The protein communication channel is uniquely characterized by its probability transition matrix. The (i, j) entry of this matrix, $\Pr(P_j \mid P_i)$, is the probability of receiving protein $P_j = (a_1^j, \ldots, a_N^j)$ given that protein $P_i = (a_1^i, \ldots, a_N^i)$ was transmitted. We assume that the protein channel is memoryless. Hence, we have

$$\Pr(P_j \mid P_i) = \prod_{k=1}^{N} \Pr(a_k^j \mid a_k^i). \tag{1}$$

From the above equation, we see that it is sufficient to study the probability transition matrix, $\mathbf{Q}(k) = \{q_{i,j}(k)\}_{1 \le i,j \le 20}$, at time k, of the amino acids.

In this paper we use two different probability transition matrices: the $\mathbf{PAM_{250}}$ probability transition matrix [20] and a first-order Markov transition probability matrix, $\mathbf{P}$. The $\mathbf{PAM_{250}}$ matrix reflects the frequencies of mutations for proteins which have diverged 250% (250 mutations per 100 amino acids). These matrices were later refined by Jones et al. [21] based on a much larger dataset. $\mathbf{P}$ is constructed from the genetic code as follows: Let $\alpha(k)$ be the probability of a base interchange of any one nucleotide at time k, all interchanges being equally probable. Assuming that the 61 codons are equally

probable and from Bayes' rule, we obtain the following formula for the probability of a transition from amino acid a to amino acid $\hat{a}$:

$$\begin{aligned}(2)\qquad \Pr(\hat{a} \mid a) &= \Pr(\{c_1, \ldots, c_n\} \mid \{b_1, \ldots, b_m\}) \\ &= \frac{1}{m} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha(k)^{h(b_j, c_i)} (1 - 3\alpha(k))^{3 - h(b_j, c_i)},\end{aligned}$$

where $\{c_1, \ldots, c_n\}$ (resp. $\{b_1, \ldots, b_m\}$) are the codons of the received (resp. transmitted) amino acid 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 $\alpha(k)$. We form the Markov probability transition matrix $\mathbf{P} = \{p_{i,j}\}_{1 \le i,j \le 20}$ by ordering the amino acids alphabetically using their one-letter standard abbreviations, e.g., $p_{1,1} = \Pr(A \mid A)$. For example, using (2), the probability of receiving amino acid N given that amino acid D was transmitted is given by

$$\begin{aligned}p_{3,12} &= \Pr(N \mid D) = \tfrac{1}{2}\{\Pr(AAC \mid GAC) + \Pr(AAC \mid GAU) + \Pr(AAU \mid GAC) \\ &\qquad + \Pr(AAU \mid GAU)\} \\ &= \tfrac{1}{2}\{\alpha(1-\alpha)^2 + \alpha^2(1-\alpha) + \alpha^2(1-\alpha) + \alpha(1-\alpha)^2\} = \alpha(1-\alpha) \approx \alpha.\end{aligned}$$

The probability transition matrix $\mathbf{P}$ is given by

$$\begin{pmatrix}
1-6\alpha & 0 & \alpha/2 & \alpha/2 & 0 & \alpha & 0 & 0 & 0 & 0 & 0 & 0 & \alpha & 0 & 0 & \alpha & \alpha & \alpha & 0 & 0 \\
0 & 1-7\alpha & 0 & 0 & \alpha & \alpha & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \alpha & 2\alpha & 0 & 0 & \alpha & \alpha \\
\alpha & 0 & 1-8\alpha & 2\alpha & 0 & \alpha & \alpha & 0 & 0 & 0 & 0 & \alpha & 0 & 0 & 0 & 0 & 0 & \alpha & 0 & \alpha \\
\alpha & 0 & 2\alpha & 1-7\alpha & 0 & \alpha & 0 & 0 & \alpha & 0 & 0 & 0 & 0 & \alpha & 0 & 0 & 0 & \alpha & 0 & 0 \\
0 & \alpha & 0 & 0 & 1-8\alpha & 0 & 0 & \alpha & 0 & 3\alpha & 0 & 0 & 0 & 0 & 0 & 0 & \alpha & 0 & \alpha & 0 & \alpha \\
\alpha & \alpha/2 & \alpha/2 & \alpha/2 & 0 & 1-\frac{23}{4}\alpha & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{3}{2}\alpha & \alpha/2 & 0 & \alpha & \alpha/4 & 0 \\
0 & 0 & \alpha & 0 & 0 & 0 & 1-8\alpha & 0 & 0 & \alpha & 0 & \alpha & \alpha & 2\alpha & \alpha & 0 & 0 & 0 & 0 & \alpha \\
0 & 0 & 0 & 0 & \frac{2}{3}\alpha & 0 & 0 & 1-7\alpha & \alpha/3 & \frac{4}{3}\alpha & \alpha & \frac{2}{3}\alpha & 0 & 0 & \alpha/3 & \frac{2}{3}\alpha & \alpha & \alpha & 0 & 0 \\
0 & 0 & 0 & \alpha & 0 & 0 & 0 & \alpha/2 & 1-7\alpha & 0 & \alpha/2 & 2\alpha & 0 & \alpha & \alpha & 0 & \alpha & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & \alpha & 0 & \alpha/3 & \frac{2}{3}\alpha & 0 & 1-\frac{11}{2}\alpha & \alpha/3 & 0 & \frac{2}{3}\alpha & \alpha/3 & \frac{2}{3}\alpha & \alpha/3 & 0 & \alpha & \alpha/6 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 3\alpha & \alpha & 2\alpha & 1-9\alpha & 0 & 0 & 0 & \alpha & 0 & \alpha & \alpha & 0 & 0 \\
0 & 0 & \alpha & 0 & 0 & 0 & \alpha & \alpha & 2\alpha & 0 & 0 & 1-8\alpha & 0 & 0 & 0 & \alpha & \alpha & 0 & 0 & \alpha \\
\alpha & 0 & 0 & 0 & 0 & 0 & \alpha/2 & 0 & 0 & \alpha & 0 & 0 & 1-6\alpha & \alpha/2 & \alpha & \alpha & \alpha & 0 & 0 & 0 \\
0 & 0 & 0 & \alpha & 0 & 0 & 2\alpha & 0 & \alpha & \alpha & 0 & 0 & \alpha & 1-7\alpha & \alpha & 0 & 0 & 0 & 0 & 0 \\
0 & \alpha/3 & 0 & 0 & 0 & \alpha & \alpha/3 & \alpha/6 & \alpha/3 & \frac{2}{3}\alpha & \alpha/6 & 0 & \frac{2}{3}\alpha & \alpha/3 & 1-\frac{17}{3}\alpha & \alpha & \alpha/3 & 0 & \alpha/3 & 0 \\
\frac{2}{3}\alpha & \frac{2}{3}\alpha & 0 & 0 & \alpha/3 & \alpha/3 & 0 & \alpha/3 & 0 & \alpha/3 & 0 & \alpha/3 & \frac{2}{3}\alpha & 0 & \alpha & 1-\frac{37}{6}\alpha & \alpha & 0 & \alpha/6 & \alpha/3 \\
\alpha & 0 & 0 & 0 & 0 & 0 & 0 & \frac{3}{4}\alpha & \alpha/2 & 0 & \alpha/4 & \alpha/2 & \alpha & 0 & \alpha/2 & \frac{3}{2}\alpha & 1-6\alpha & 0 & 0 & 0 \\
\alpha & 0 & \alpha/2 & \alpha/2 & \alpha/2 & \alpha & 0 & \frac{1}{4}\alpha & 0 & \frac{3}{2}\alpha & \alpha/4 & 0 & 0 & 0 & 0 & 0 & 0 & 1-6\alpha & 0 & 0 \\
0 & 2\alpha & 0 & 0 & 0 & \alpha & 0 & 0 & 0 & \alpha & 0 & 0 & 0 & 0 & 2\alpha & \alpha & 0 & 0 & 1-7\alpha & 0 \\
0 & \alpha & \alpha & 0 & \alpha & 0 & \alpha & 0 & 0 & 0 & 0 & \alpha & 0 & 0 & 0 & \alpha & 0 & 0 & 0 & 1-6\alpha
\end{pmatrix}$$

For display clarity, we omitted the dependence of the point mutation rate $\alpha(k)$ on the time k. Observe that $\mathbf{P}$ takes into account all possible mutations between amino acids whether they are accepted or rejected by natural selection whereas the $\mathbf{PAM_{250}}$ transition matrix is estimated from phylogenetic trees of protein sequences and hence takes into account the accepted mutations only.

Let $\mathbf{p}_0$ be the row probability vector of the initial distribution of the amino acids (at time 0). It is straightforward to show that the row probability vector of the amino acids at time k, $\mathbf{p}_k$, is given by

$$(3)\qquad \mathbf{p}_k = \mathbf{p}_0 \mathbf{Q}(1)\mathbf{Q}(2)\cdots\mathbf{Q}(k).$$

2.1. *Constant Point Mutation Rate.* In this subsection we assume that the point mutation rate is constant over time, i.e., $\alpha(k) = \alpha$, for all times $k \geq 0$. Equation (3) then becomes

$$\mathbf{p}_k = \mathbf{p}_0 \mathbf{Q}^k. \tag{4}$$

PROPOSITION 1. *Consider an initial probability distribution of the amino acids at time* 0, $\mathbf{p}_0$. *Then the probability distribution of the amino acids converges, over time, towards the stationary distribution given by*

$$\begin{cases} \mathbf{s_1}, & \text{if } \mathbf{Q} = \mathbf{P}; \\ \mathbf{s_2}, & \text{if } \mathbf{Q} = \mathbf{PAM_{250}}, \end{cases}$$

where

$$\mathbf{s_1} = \left(\tfrac{4}{61}, \tfrac{2}{61}, \tfrac{2}{61}, \tfrac{2}{61}, \tfrac{2}{61}, \tfrac{4}{61}, \tfrac{2}{61}, \tfrac{3}{61}, \tfrac{2}{61}, \tfrac{6}{61}, \tfrac{1}{61}, \tfrac{2}{61}, \tfrac{4}{61}, \tfrac{2}{61}, \tfrac{6}{61}, \tfrac{6}{61}, \tfrac{4}{61}, \tfrac{4}{61}, \tfrac{1}{61}, \tfrac{2}{61}\right) \tag{5}$$

and

$$\begin{aligned} \mathbf{s_2} = (&0.0873, 0.0338, 0.0479, 0.05, 0.0383, 0.0909, 0.0330, 0.0375, \\ &0.0808, 0.0844, 0.0143, 0.0411, 0.0522, 0.0390, 0.0406, 0.0704, \\ &0.0594, 0.0651, 0.0075, 0.0294). \end{aligned} \tag{6}$$

In order to make biological sense of the limiting distribution vectors $\mathbf{s_1}$ and $\mathbf{s_2}$, we compare them with the experimental distribution of amino acids computed in the literature [22], [21], [23], [24]. We found that there are some fluctuations between the different experimental distributions. The reason behind this disparity is that different experiments use different sets of organisms and different protein families. Let us denote by $\mathbf{r}$ the experimental probability vector of the amino acids. Table 1 displays the correlation coefficients between the different experimental distributions and the limiting distributions $\mathbf{s_1}$ and $\mathbf{s_2}$. Since $\mathbf{PAM_{250}}$ estimates the rate of accepted mutations only, we find that the limiting distribution $\mathbf{s_2}$ has a higher correlation with the experimental distribution $\mathbf{r}$ than the limiting distribution $\mathbf{s_1}$. Moreover, the highest correlation was obtained between $\mathbf{s_2}$ and the experimental distribution computed in [23]. Figure 3 shows the plot of $\mathbf{r}$ in [23]

Table 1. Correlation between the Experimental Frequencies of Amino Acids and the Limiting Distributions $\mathbf{s_1}$ and $\mathbf{s_2}$.

Experimental distribution $\mathbf{r}$	Correlation coefficient between $\mathbf{r}$ and $\mathbf{s_2}$	Correlation coefficient between $\mathbf{r}$ and $\mathbf{s_1}$
$\mathbf{r}$ in [23]	0.96	0.66
$\mathbf{r}$ in [22]	0.937	0.632
$\mathbf{r}$ in [24] Eukaryotes	0.824	0.74
$\mathbf{r}$ in [24] Bacteria	0.836	0.701
$\mathbf{r}$ in [24] Archaea	0.76	0.602
$\mathbf{r}$ in [24] all taxa	0.834	0.7

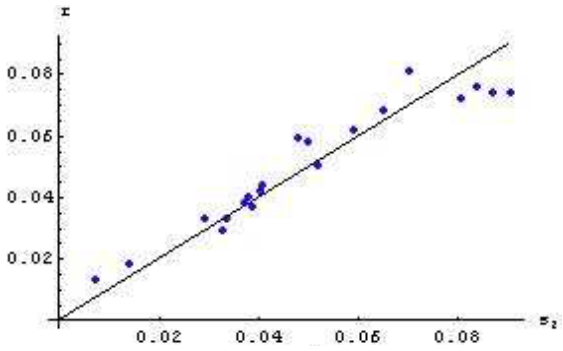

Fig. 3. The experimental distribution of amino acids **r** in [23] vs. the limiting distribution $\mathbf{s_2}$ given by the $\mathbf{PAM_{250}}$ probability transition matrix.

versus $\mathbf{s_2}$. Another interesting observation is that the limiting distributions have higher correlations with the experimental frequencies of amino acids calculated from eukaryotes and bacteria than the experimental frequency calculated from archaea (see rows 3–5 of Table 1).

Notice that $\mathbf{s_1}$ is proportional to the number of codon assignments for the amino acids. So, $\mathbf{s_1}$ is the distribution of the amino acids if the codons were randomly distributed in the genome. Equivalently, we can view $\mathbf{s_1}$ as the distribution of amino acids if all randomly distributed point mutations were accepted by Nature (i.e., survived). According to this view, the discrepancy between $\mathbf{s_1}$ and $\mathbf{s_2}$ can be related to the relative probability of survival of the amino acids after mutations. We shall divide the amino acids into classes C_1, C_2, C_3, C_4 and C_6 , the subscripts indicating the number of codons for each class. For example, the class C_1 contains two amino acids: Met (M) and Trp (W), i.e., $C_1 = \{M, W\}$. The mean experimental in [23] and limiting distributions using both matrices **P** and $\mathbf{PAM_{250}}$ are displayed in Table 2. The mean experimental and limiting distributions, for each class, are very close except for the class of amino acids corresponding to six codons obtained from the limiting distribution using the probability transition matrix **P**. The reason is that arginine, which is coded by six codons, appears with a much lower frequency than $\frac{6}{61}$. This has been ascribed to the rare appearance of the CG base doublet so that, in fact, in most observed proteins, arginine is coded only by AGA and AGG [4].

Table 2. Mean Experimental and Limiting Distributions of the Amino Acid Classes.

Classes	Mean experimental probability in [23]	Mean limiting probability (**P**)	Mean limiting probability (PAM)
C_1	0.0155	0.0163	0.01075
C_2	0.045	0.0327	0.044
C_3	0.04843	0.0492	0.0508
C_4	0.0656	0.0656	0.0709
C_6	0.0663	0.0983	0.0665

A question naturally arises now: what is the rate of convergence and how is this rate related to the rate of point mutation α? The answer is provided in the following proposition:

PROPOSITION 2. *The sequence of probability vectors* $\{\mathbf{p_k}\}_{k\geq 1}$ *converges at a geometric rate with parameter* $|\lambda_2|$, *where*

$$\begin{cases} |\lambda_2| = 0.53, & \text{if} \quad \mathbf{Q} = \mathbf{PAM_{250}}; \\ |\lambda_2| \leq 1 - \frac{1}{2}\alpha, & \text{if} \quad \mathbf{Q} = \mathbf{P}. \end{cases} \tag{7}$$

Thus, the convergence rate for $\mathbf{P}$ is no slower than $\mathsf{O}((1 - \frac{1}{2}\alpha)^k)$. Moreover, when α decreases, the convergence is slower and vice versa. This result is somehow intuitive and ascertains that no evolution is possible if the point mutation rate $\alpha = 0$.

2.2. *Time-Varying Point Mutation Rate*. In this section we consider a rate of point mutation, $\alpha(k)$, which varies in time. Consider the products $\mathbf{T}_{p,k} = \{t_{i,j}^{(p,k)}\} = \mathbf{Q}_{p+1}\mathbf{Q}_{p+2}\cdots \mathbf{Q}_{p+k}$ for every $p \geq 0$. For a fixed p, let t be the smallest integer satisfying $\mathbf{T}_{p,t} > 0$, in the sense that all its entries are strictly positive.

DEFINITION 1 (Weak and Strong Ergodicity) [25]. The forward products $\mathbf{T}_{p,k}$ are said to be *weakly ergodic* if

$$t_{i,s}^{p,k} - t_{j,s}^{p,k} \xrightarrow{k\to\infty} 0 \qquad \text{for each } i,\ j,\ s,\ p. \tag{8}$$

If weak ergodicity is obtained and the $t_{i,s}^{p,k}$ themselves tend to a limit for all i, s, p, i.e., $t_{i,j}^{(p,k)} \xrightarrow{k\to\infty} v_j^{(p)}$, then we say *strong ergodicity* is obtained.

Moreover, if strong ergodicity is obtained, then the limit row vector $\mathbf{v}_p = \{v_j^{(p)}\}$ is a probability vector and is independent of $p \geq 0$, i.e., $\mathbf{v}_p = \mathbf{v}$ [25]. Hence, strong ergodicity is equivalent to the existence of the limit of $\mathbf{T}_{p,k}$ as $k \to \infty$, for all $p \geq 0$.

DEFINITION 2 [25]. A matrix $\mathbf{Q} = \{q_{i,j}\}$ is called a *scrambling* matrix if given any two rows β and δ, there is at least one column ρ such that $q_{\beta,\rho} > 0$ and $q_{\delta,\rho} > 0$.

It is easy to show that if every matrix $\mathbf{Q}(k)$ is scrambling, then so is $\mathbf{T}_{p,k}$, $p \geq 0$.

THEOREM 1 (Weak Ergodicity Result). *Consider a finite number of PAM matrices denoted by* $\mathbf{PAM(1)}, \cdots, \mathbf{PAM(N)}$, *where* $\mathbf{PAM(i)}$ *can be* $\mathbf{PAM_1}$ *or* $\mathbf{PAM_{160}}$ *or* $\mathbf{PAM_{250}}$, *etc., for all* $i = 1, \ldots, N$. *Consider the sequence*: $\mathbf{T}_{p,k} = \mathbf{t}_{p+1}\mathbf{t}_{p+2}\cdots\mathbf{t}_{p+k}$, *where each* $\mathbf{t}_i \in \{\mathbf{PAM(1)}, \ldots, \mathbf{PAM(N)}\}$. *That is at each time* k, *the probability transition matrix is some PAM matrix* (*the evolutionary time of the PAM matrix and the time* k *are not necessarily the same*). *Then* $\mathbf{T}_{p,k}$ *is weakly ergodic at a uniform geometric rate for all* $p \geq 0$. *Consequently, the sequence* $\{\mathbf{p}_k\}_{k\geq 1}$, *in* (3), *tends to a sequence of distributions independently of* $\mathbf{p_0}$.

If we approximate the matrices $\mathbf{PAM_k}$ by $\mathbf{PAM_1^k}$, the sequence

$$\mathbf{T}_{p,k} = \mathbf{PAM^{p+1}PAM^{p+2}} \cdots \mathbf{PAM^{p+k}}$$

becomes strongly ergodic. In particular, the sequence $\{\mathbf{p}_k\}_{k\geq 1}$, in (3), converges to the limiting distribution $\mathbf{s_2}$ given in (6).

THEOREM 2 (Strong Ergodicity Result). *Consider a point mutation rate, $\alpha(k)$, which is bounded uniformly on k, i.e., $0 < a \leq \alpha(k) \leq b < 1$, for some $a > 0$ and $b < 1$. Then the products $\mathbf{T}_{p,k} = \mathbf{P}_{p+1} \cdots \mathbf{P}_{p+k}$ are strongly ergodic. Thus, the sequence $\{\mathbf{p}_k\}_{k\geq 1}$, in (3), converges towards the stationary distribution $\mathbf{s_1}$, in (5), independently of the initial distribution $\mathbf{p}_0$. Moreover, the convergence rate is at least geometric with parameter $(1-\gamma^t)^{1/t}$, where $\gamma = \min\{a/6, 1-9b\}$.*

The time-varying point mutation rate analysis shows that, even in the more realistic scenario of a time-varying channel, the distribution of the amino acids can converge to a fixed limiting distribution (Theorem 2) independently of the initial distribution of amino acids. In particular, this result implies that, after a sufficiently long time, the channel characteristics will determine the final distribution which will be independent of the initial distribution. This conclusion has very different ramifications on bioinformatics than on communication engineering: The convergence analysis in engineering is interpreted as a loss of information after an infinite number of transmissions. The reason is that, in communications, only the initial distribution (i.e., the message) is used to convey information and not the channel. In bioinformatics, on the other hand, the final distribution of amino acids captures the information of the channel (i.e., the mutations) regardless of the initial distribution. The critical information in modelling the channel of evolution is therefore the representation of the channel and not the starting point of the evolutionary process. These implications are verified experimentally [22].

3. Genomic Structure

3.1. *Proposed Role of Introns*. We propose that introns maintain a genius balance between stability and adaptability in eukaryotic genomes as follows:

1. Introns reduce the probability of mutation error in the coding regions (i.e., exons) by serving as decoys which absorb isolated mutations. According to this view, introns protect coding regions in the DNA sequence from frequent errors in the same way that hollow uninhabited structures are used by the military to protect important installations, such as aircraft hangars and missile launching facilities, from a bomb attack by serving as a "dummy" target that resembles the protected structure. It is important to emphasize that the role of introns is not to ensure a perfect (errorless) communication system, but to temper the effervescence of the ever-changing genome under the chemical, physical and environmental conditions. Perfect information transmission will spell stagnation and ultimately extinction. This is the major difference between an engineering communication system and the biological communication system.

2. Introns drive biological evolution by increasing the rate of recombination of exons and consequently participate in the creation of new genes [26], [27]. To understand the role of introns in the assembly of new genes, we found no better explanation than Gilbert's statement: "Consider a new gene made by a new combination of regions of earlier genes by an unequal crossover, a rare event at the DNA level, that matches small, similar sequences between two DNAs. To make a new protein that contains the first part of one protein with the second part of another requires such a rare, and in frame, event. However, if the regions that encode parts of the protein are separated by 1,000–10,000 base long introns along the DNA, a process of unequal crossing-over occurring anywhere within that intron between the exons will create a new combination of exons" [16]. Recently, it has been experimentally proven that intron lengths are negatively correlated with the rate of recombination in Drosophila Melanogaster and humans [28]. That is, the advantage of longer introns is expected to decrease inversely with the rate of recombination. Hence, in the chromosomal regions where crossing over is infrequent, introns tend to be larger to increase the rate of recombination between exons. Whether introns were used to assemble the first genes or not is not relevant to our investigation as long as we have biological evidence that new genes were and are currently created through the mechanism of unequal crossover.

The proposed dual role of introns serves to provide a balance between two competing biological evolutionary functions: stability and adaptability. In the remainder of this paper, we prove the stability role of introns. However, first we address the question of the need for a decoy for mutations besides the well-known proof-reading mechanisms.

3.2. *Why Does Nature Need a Decoy for Mutations Besides the Proof-Reading Mechanisms*? DNA repair mechanisms are constantly operating in cells. In human cells both normal metabolic activities and environmental factors can result in as many as a million molecular lesions per cell each day. Consequently, DNA repair mechanisms are essential for the survival of the organism. However, it is also known that these DNA repair mechanisms are not 100% efficient and many errors remain undetected or uncorrected in the genome. Let us anecdotically compare the efficiency of the Reed–Solomon code and the genetic error correction mechanisms. The potential efficiency of a code is a function of the number of redundant bits. A commonly used Reed–Solomon code, in CD players for instance, uses a codeword length of 255 bytes, of which 223 bytes are data and 32 bytes are parity. Thus, the redundancy rate of the Reed–Solomon code is $\frac{32}{255} = 13\%$. The human genome contains about 30,000 genes, of which about 130 code for DNA repair enzymes [18]. Assuming that the genes have roughly the same number of nucleotides, the redundancy rate of the human error correction mechanism is $130/30{,}000 = 0.43\%$! Hence, despite all the excitement that the discovery of DNA repair mechanisms brought (especially to creationists), this simple argument indicates that the repair mechanism of the human genome, for instance, is unlikely to be very efficient. We argue does Nature uses introns as a decoy for mutations to achieve a lower error rate? However, one can legitimately ask: Why wouldn't nature invest in more error correction mechanisms rather than carry this enormous decoy luggage? Several reasons lie behind this choice: First, if nature had to design error correction codes to control the exact rate of mutation required

to maintain life and simultaneously encourage evolution, it would need to know the exact distribution and form of all possible mutations which occurred in the past and will occur in the future. Designing complex error correcting codes for a given noise model might be completely useless in the face of dynamic noise characteristics. Second, a reduction in the error rate comes at the price of an increase in complexity. Nature might have preferred to spend more energy in carrying the decoy sequences rather than investing in complex and costly error repair enzymes.

3.3. *Genomic Structure*: *Deterministic Analysis*

PROPOSITION 3. *Consider a genome of length T. Assume that the point mutation rate is randomly distributed in the genome, i.e., the occurrence of mutations is independent and identically distributed in all regions of the genome. Then the probability of error is a decreasing function of the length of introns and is independent of the distribution of introns in the genome.*

Hence, we see that a binomial error model does not account for the biological exon (or intron) length distribution inside the genome. In other words, the biological intron–exon distribution would be equivalent, from an error robustness criterion, to the distribution which groups all exons in the beginning of the gene and all introns at its end. Therefore, we need to consider a different mutation model, which can account for the observed intron–exon structure in eukaryotic genomes. We propose a Poisson mutation model. This choice is justified by numerous arguments. First, the Poisson distribution is the limiting distribution of the binomial when the probability of error is small and the genome size is large such that the rate of point mutation in a unit interval is held constant (De Moivre–Laplace theorem [29]). 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. In the remainder of this paper we assume that the mutations are Poisson distributed in the genome.

Assume now that there are K exons of total length M in a gene of T nucleotides. Let l_k be the length of exon k. In the following proposition we answer the question: "What are the optimal exon lengths, l_k^*, $k = 1, \ldots, K$, which minimize the probability of error in the gene?"

PROPOSITION 4. *Assume that the mutations are Poisson distributed with rate λ. Consider a genome of length T nucleotides including K exons having total length M. Let l_k be the length of the k*th *exon. Then the probability of error is given by*

$$P_e = 1 - e^{-\lambda K T} \prod_{k=1}^{K} \sum_{n=0}^{T-l_k} \frac{\lambda^n (T - l_k)^n}{n!}. \tag{9}$$

Since $l_k \leq M$ for all $k = 1, \ldots, K$, we obtain an upper bound on the probability of error by truncating the summation in (9) to $T - M$ instead of $T - l_k$. Minimizing the maximum probability of error, $P_e^{\max}$, is more tractable analytically than minimizing the probability of error in (9). Using the Lagrange multiplier technique, with constraint $\sum_{k=1}^{K} l_k = M$, and taking the derivative of $P_e^{\max}$ with respect to l_k, we obtain the following

coupled system for the optimal exon lengths:

$$l_{i_0} = M \frac{[\prod_{k\neq i_0} \sum_{n=0}^{T-M} (\lambda^n (T-l_k)^n/n!)][\sum_{n=1}^{T-M} \lambda^n (T-l_{i_0})^{n-1}/(n-1)!]}{\sum_{j=1}^{K} [\prod_{k\neq j} \sum_{n=0}^{T-M} (\lambda^n (T-l_k)^n/n!)][\sum_{n=1}^{T-M} \lambda^n (T-l_j)^{n-1}/(n-1)!]}. \tag{10}$$

An obvious solution to the system in (10) is obtained when $l_k^* = M/K$ for all $k = 1, \ldots, K$. This surprising simple result states that the optimal exon lengths are distributed according to a delta function centered at the mean value M/K. However, in Nature the exon lengths are not uniformly distributed in the genome (see Figure 4). The reason this deterministic analysis fails in capturing the intron–exon distribution is that the genome is not a deterministic entity but rather a continuously evolving one. Therefore, a stochastic model for the exon lengths would be more appropriate to describe the genome's dynamic nature correctly. The deterministic analysis does, however, capture some characteristics of the biological data in the following sense:

PROPOSITION 5. *Let $\delta_{M/K}$ be the delta function centered at M/K. For every $\rho > 0$, consider the measure d_ρ between a continuous unimodal probability density function f_X and $\delta_{M/K}$ given by*

$$d_\rho(\delta_{M/K}, f_X) = 1 - Pr\left(X \in \left[\frac{M}{K} - \rho, \frac{M}{K} + \rho\right]\right). \tag{11}$$

Let x_0 be the mode of f_X. Then $\operatorname{argmin}_{x_0} d_\rho = M/K$. *That is the mode of f_X, which minimizes the measure d_ρ, is equal to M/K.*

The biological exon distribution is asymmetric given that its support is $[0, \infty]$. The mode of asymmetric distributions is always less than or equal to their mean. From Proposition 5, the distribution, which best approximates $\delta_{M/K}$ in the d_ρ measure sense, would have its mode very close to its mean. Amazingly, the exon length distribution of the human genome has its mode almost equal to its mean obtained at about 170 nucleotides (see Figure 4)!

Even though the deterministic analysis gave some insights on the optimality of the biological exon length distribution from an error minimization criterion, a stochastic model for the exon distribution is needed to capture the dynamics of the evolving genome.

3.4. *Genomic Structure*: *Stochastic Analysis*. In this subsection we re-address the probability of error optimization problem formulated above assuming a stochastic distribution of the exon lengths. The following proposition establishes the new expression for the probability of error assuming an infinite genome length, i.e., $T = \infty$.

PROPOSITION 6. *Assume that there are K exons in a genome infinitely long. Let $p(l)$ be the continuous density of the exon lengths. Assume that the mutations are Poisson distributed with parameter λ. Then the probability of error is given by*

$$P_e = 1 - \left(\int_0^\infty e^{-\lambda l} p(l)\, dl\right)^K. \tag{12}$$

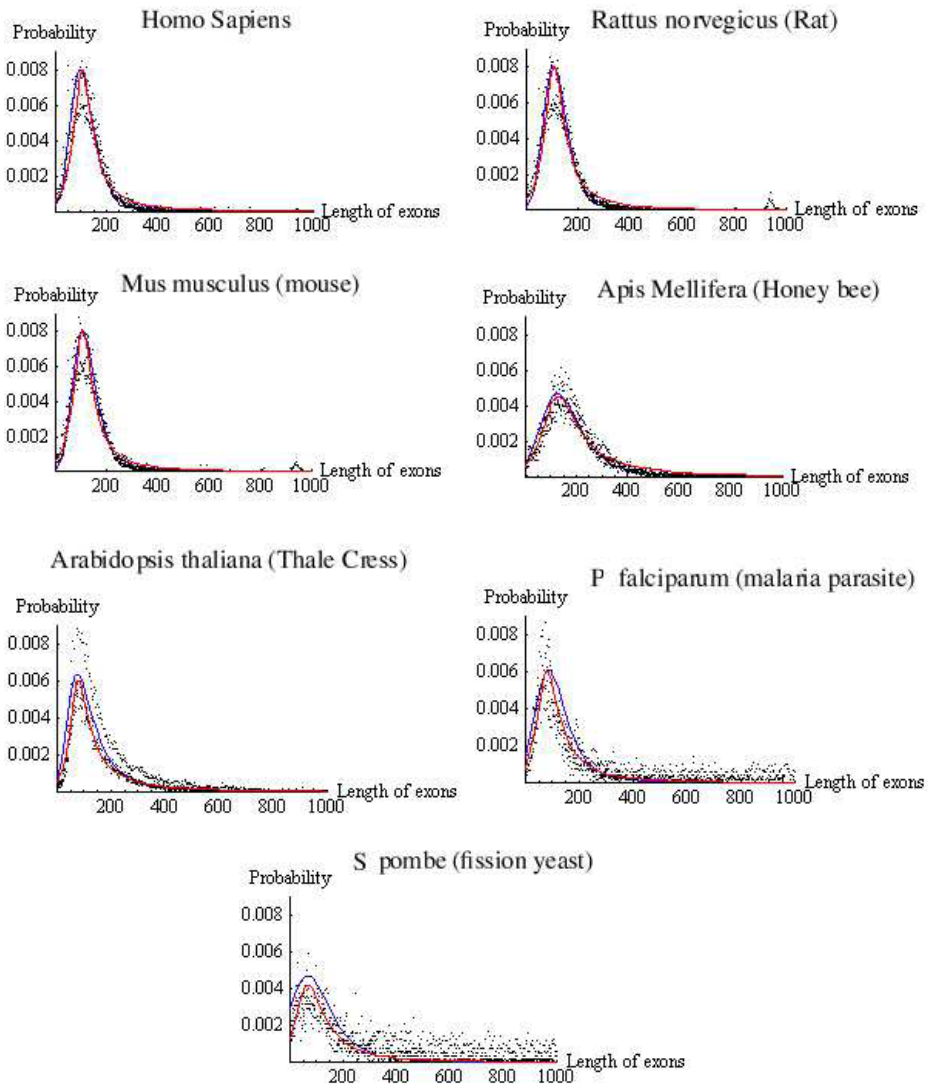

Fig. 4. Exon length distribution: the data points represent the biological data; the red curve is the optimal density, which minimizes the probability of error; and the blue curve is the fitted α-stable distribution. The graphs of the densities are truncated at exon lengths of 1000 nucleotides.

We want to determine the optimal exon length distribution, $p^*(l)$, which minimizes the probability of error subject to $\int_0^\infty p^*(l)\, dl = 1$. It can be easily shown that the delta function centered at 0, δ_0, satisfies this optimization problem. This solution is somehow intuitive: no exons implies no error! In order to get a meaningful solution to this optimization problem, we need to impose more constraints on the exon length distribution. For instance, the mean exon length should be larger than a pre-specified number l_0 or, more generally, the αth moment of $p(l)$ should be larger than l_0. Consequently, the stochastic

optimization problem is reformulated as follows:

$$p^*(l) = \operatorname*{argmax}_{p(l)} \int_0^\infty e^{-\lambda l} p(l)\, dl, \tag{13}$$

$$\text{subject to} \quad (1) \quad \int_0^\infty p(l)\, dl = 1;$$

$$(2) \quad \int_0^\infty l^{1+\alpha} p(l)\, dl \geq l_0, \qquad \text{for some} \quad \alpha \geq 0.$$

The optimization problem formulated in (13) is solved using the Euler–Lagrange equation. We obtain

$$p^*(l) = \frac{p_0(1+\mu)}{e^{-\lambda l} + \gamma l^{1+\alpha} + \mu}, \tag{14}$$

where μ and γ are the Lagrange multipliers, which are determined numerically. The parameter α determines the tail decay of the distribution. Taking the derivative of p^*, it is easy to show that it has a unique maximum. Observe that the $(1+\alpha)$th moment of $p(l)$ is infinite; thus satisfying condition (2) in (13). This infinite moment agrees with the heavy tail characteristic of the biological exon length distribution (see Figure 4).

Having determined analytically the optimal exon length distribution, it is interesting to ask ourselves: "How can Nature generate such a distribution? Is there a simple enough model for exon generation, which leads to the distribution p^*?" The answer is investigated in the next subsection.

3.5. *A Diffusive Random Walk Model.* The distributions of initial, internal and terminal exons in various organisms were shown to be different [30]. However, our purpose is to model the overall distribution of all exons in eukaryotic genomes. Interestingly, prior experimental results show that this distribution is identical except for translation and scale parameters (see Figure 4). Exon shuffling models [31] and insertion and deletion of nucleotides have been confirmed biologically for many primitive and higher-order organisms [32]. We propose to model the formation of an exon by concatenation, insertion and deletion of sub-exons (of different lengths). If exons were formed by insertion and deletion mechanisms, their lengths would follow some kind of random walk. The length of the exon at any given time corresponds to the position of the random walk. After N steps, the length of the exon, X_N, is the sum of N random displacements, i.e.,

$$X_N = \sum_{i=1}^{N} l_i. \tag{15}$$

We are interested in the limiting distribution of X_N as $N \to \infty$. The experimental analysis of DNA sequences has shown that non-coding DNA exhibits long-range dependence whereas coding DNA behaves more like a random sequence [33]–[35]. Therefore, we assume that the sub-exons are formed independently by a stochastic process according to a distribution $f(l)$. Given the heavy tail characteristic of the empirical exon length distribution (see Figure 4), we assume that $f(l)$ is power law distribution, i.e.,

$$f(l) = \begin{cases} 0, & l < l_0, \\ Al^{-(\alpha+1)}, & l \geq l_0, \end{cases} \tag{16}$$

where $0 < \alpha < 2$ and l_0 is a cutoff at short lengths to allow the function to be normalizable; the normalization constant is $A = \alpha l_0^{\alpha}$. By the Generalized Central Limit Theorem [36], the density of X_N tends towards an α-stable distribution $S_\alpha(x \mid \beta, \sigma, \xi)$. Since Paul Levy found the class of α-stable distributions, in 1925, as simple exceptions to the Central Limit Theorem, a vast amount of knowledge has been accumulated about the properties of these probability distributions, especially infinite moments, elegant scaling properties and the inherent self-similarity property. They have been found to provide useful models in the study of physical and economic systems, especially phenomena with large fluctuations and high variability that are not compatible with the Gaussian models. Except the Gaussian, the Cauchy and the Levy distributions which are special cases of the stable class, there is no exact expression of the probability density function of an α-stable distribution. α-Stable distributions are defined by their characteristic function. Four parameters are needed: an index of stability $\alpha \in (0, 2]$, a skewness parameter $\beta \in [-1, 1]$, a scale parameter $\sigma > 0$ and a location parameter $\xi \in (-\infty, +\infty)$. There are multiple parametrizations of α-stable distributions. For numerical purpose, we use a variant of the M-parametrization of Zolotarev [37] with the following characteristic function [38]:

$$\exp^{\iota\omega X} = \begin{cases} \exp^{-\sigma^{\alpha}|\omega^{\alpha}|[1+\iota\beta\tan(\alpha\pi/2)\operatorname{sign}(\omega)((\sigma|\omega|)^{1-\alpha}-1))+\iota\xi\omega]}, & \text{if} \quad \alpha \neq 1; \\ \exp^{\sigma|\omega|[1+\iota\beta(2/\pi)\operatorname{sign}(\omega)\ln(\sigma|\omega|)]+\iota\xi\omega}, & \text{if} \quad \alpha = 1. \end{cases} \tag{17}$$

The above parametrization is a scale and location family of distributions: if $Y \sim S_\alpha(\beta, \sigma, \xi)$, then for any a, b, $aY + b \sim S_\alpha(\operatorname{sign}(a)\beta, |a|\sigma, a\xi + b)$. Other related issues of stable distributions are discussed in [38]. Some of the prominent properties of α-stable distributions are: heavy tail, skewness (when $\beta \neq 0$) and smooth unimodal density. Their asymptotic behavior is described by

$$\lim_{|x|\to\infty} S_\alpha(x \mid \beta, \sigma, \xi) = \frac{C}{|x|^{1+\alpha}}, \tag{18}$$

where C is some constant [36]. Hence, from (14), we see that the optimal distribution p^* is asymptotically equivalent to an α-stable distribution. Nature would prefer to generate a simple random walk rather than solve the Euler–Lagrange equation!

4. Experimental Results. All exon lengths for each of the Homo sapiens (Human), Rattus norvegicus (Rat), Mus musculus (mouse), Apis mellifera (Honey bee), Schizosaccharomyces pombe (fission yeast), Plasmodium falciparum (malarial parasite) and Arabidopsis thaliana (thale cress) genomes were studied. The data files used were obtained from the NCBI web site: "ftp://ftp.ncbi.nih.gov/genomes". Exons tagged as CDS were included in the analysis. The NCBI handbook makes clear that CDS refers to the portion of a genomic DNA sequence that is translated. Alternative spliced variants were kept in the data, so some exons can be recorded several times from a given gene.

An initial data analysis is presented in Table 3. Of the seven different organisms examined, H sapiens contained the greatest number of exons, 281,975. S pombe has the least number of exons of the organisms analyzed here. The descriptive statistics for H sapiens, M musculus, R norvegicus, A mellifera and A thaliana are similar. The two

Table 3. Descriptive Statistics of Exon Lengths for the Seven Organisms.

	Nb exons	Mean	Stdev	Min	Max
H sapiens	281,975	167	233	1	17,105
R norvegicus	185,769	177	378	1	9,820
M musculus	226,498	178	326	1	16,625
A mellifera	32,753	234	320	1	7,241
S pombe	9,772	698	1,038	1	11,099
P falciparum	12,660	943	1,957	2	27,815
A thaliana	164,986	228	722	1	6,040

single cellular organisms, S pombe and P falciparum, have considerably higher average exon lengths as well as greater exon length variation than all the other organisms. For all the organisms the mean exon length is greater than the median exon length, indicating a right-skewed distribution. Figure 4 shows the biological data, the optimal density and the α-stable distribution of the analyzed organisms. For α-stable density fitting, we used the Mathematica package for stable distributions available from J. P. Nolan's website: "academic2.american.edu/∼jpnolan". The parameter α was estimated by plotting the data on a log-log scale and estimating the slope: If we order the data $X(1) \geq X(2) \geq \cdots \geq X(n)$ (the order statistics of the empirical data) then we can estimate $y = P(X > t)$ by taking $y = i/n$ and $t = X(i)$. A plot of the points $(t, y) = (\ln(X(i)), \ln(i/n))$ should fit a straight line with slope $-\alpha$. Figure 5 shows a least mean square fitting of the tail of the human exons empirical distribution for a 1.5-stable distribution and the Cauchy distribution, which corresponds to $\alpha = 1$. The stable distributions $S_{1.5}(l \mid 0.9, 35, 135)$, $S_{1.5}(l \mid 0.85, 35, 140)$, $S_{1.5}(l \mid 0.9, 60, 190)$, $S_{1.5}(l \mid 0.9, 35, 143)$, $S_{1.5}(l \mid 0.9, 60, 130)$, $S_{1.5}(l \mid 0.9, 46, 135)$ and $S_1(l \mid 0.85, 45, 332)$ fit the exon length distributions of H sapiens, R Norvegicus, A Mellifera, M musculus, S pombe, P falciparum and A thaliana, respectively. The same α was used to display the optimal density $p^*(l)$ for these organisms. The mutation rate λ can be interpreted as the average rate of accepted mutations since the beginning of life on Earth.

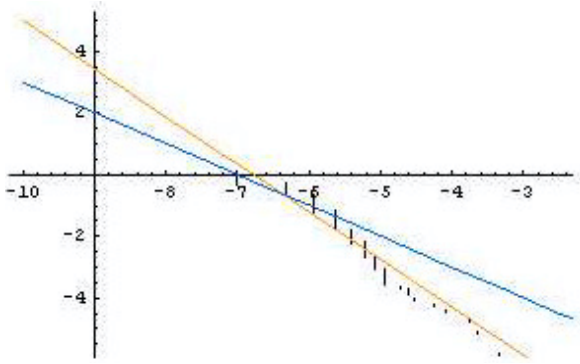

Fig. 5. The log–log plot of the human exons empirical data. The slope of the blue line is equal to -1 and corresponds to fitting a Cauchy distribution to the data. The slope of the orange line is equal to -1.5, and corresponds to an α-stable fitting of the data.

The experimental results in Figure 4 show that the empirical data of the exon length distribution becomes noisier when we go down in the scale of evolution (i.e., for more primitive organisms) and the fit becomes less accurate. This seems to be in accordance with our claim about the stability role of introns, since primitive eukaryotic organisms have few introns and rely on mutations for diversification and adaptation. It would be very interesting, however, to investigate deeper the relative roles of mutations and crossover in different organisms to ensure stability and adaptation.

5. Conclusions and Future Work. We proposed a communication channel between protein sets to model the transmission of genetic information during cell division or asexual reproduction. The encoder of the protein communication channel does not exist biologically. Nature cleverly circumvented the encoding process by ensuring that organisms contain both proteins and the encoded message: the DNA. By decoding DNA into proteins, organisms come into being. A concatenation in-series of time-dependent protein communication channels represents a channel through time: the channel of evolution. Using the $\mathbf{PAM_{250}}$ probability transition matrix and a Markov probability transition matrix constructed from the genetic code, we investigated the dynamics of the channel of evolution in both cases of constant and time-varying point mutation rates. Specifically, we showed that the distribution of the amino acids converges, at a geometric rate, towards a fixed distribution independently of the initial abundance of the amino acids. This convergence result ascertains that a parent organism will be unrelated to its offspring of many generations, no matter how small the mutation rate is as long as it is non-zero. We can obtain similar results with other amino acid substitution matrices, e.g., the BLOSUM [39] probability transition matrix constructed from the log-odds BLOSUM matrix. The mathematical tools used will apply, under some mild conditions, on the matrices.

In the second part of the paper we investigated the structure of the genetic codeword, the DNA. we proved that the introns play the role of a decoy for mutations. It is important to emphasize that the role of introns is not to ensure a perfect (errorless) communication system, but to temper the effervescence of the ever-changing genome under the chemical, physical and environmental conditions. Perfect information transmission will spell stagnation and ultimately extinction. This is the major difference between an engineering communication system and the biological communication system. We also maintain that introns increase the rate of evolutionary adaptation by providing hot spots for genetic recombination. The proposed dual role of introns serves to provide a balance between stability and adaptability. It is interesting to note that the role of introns in protection against mutations is enhanced by increasing the size of the intron regions. On the other hand, the function of introns in encouraging recombination depends on the presence of long contiguous nucleotide sequences in introns. In order to moderate the adaptability rate of the genomic sequence, the length of contiguous nucleotide sequences must be limited. Indeed, most eukaryotes display multiple intron regions within a single gene. Introns therefore seem to control the balance between stability and adaptability of the genomic sequence. One important consequence of this interpretation is that primitive eukaryotes, which have few or no introns at all, rely on mutations for diversification and evolution; whereas higher-order organisms, which have many introns, are more stable with respect to mutations but can undergo big jumps in evolution due to formation of new

genes via unequal crossover in intron regions. The proper interpretation of our model is therefore consistent with both prokaryotes and eukaryotes and can be used to help our understanding of the mystery of missing links in evolution.

Appendix

PROOF OF PROPOSITION 1. The probability transition matrices $\mathbf{P}$ and $\mathbf{PAM_{250}}$ are irreducible and aperiodic. Therefore, from the Perron–Frobenius theorem [40], there exists a unique stationary probability row vector $\mathbf{s_1}$ (resp. $\mathbf{s_2}$) such that the sequence of powers $\{\mathbf{p}_0\mathbf{P}^k\}_{k\in\mathbb{N}}$ (resp. $\{\mathbf{p}_0\,\mathbf{PAM_{250}}^k\}_{k\in\mathbb{N}}$) approaches the fixed probability vector $\mathbf{s_1}$ (resp. $\mathbf{s_2}$) as $k \to \infty$. Moreover, $\mathbf{s_1}$ and $\mathbf{s_2}$ are independent of the initial distribution $\mathbf{p}_0$. The stationary probability vector $\mathbf{s_1}$ (resp. $\mathbf{s_2}$) is the unique eigenvector of the matrix $\mathbf{P}$ (resp. $\mathbf{PAM_{250}}$), corresponding to the eigenvalue 1 and such that $\mathbf{s_1 1} = 1$ (resp. $\mathbf{s_2 1} = 1$), where $\mathbf{1}$ is the column vector with all its entries equal to 1. □

PROOF OF OF PROPOSITION 2. The matrix $\mathbf{Q} \in \{\mathbf{P}, \mathbf{PAM_{250}}\}$ is an irreducible, aperiodic and stochastic matrix. Therefore, the eigenvalues of $\mathbf{Q}$ can be ordered by $1 > |\lambda_2| \geq \cdots \geq |\lambda_t|$. As $k \to \infty$, $\mathbf{Q}^k = \mathbf{Q}_\infty + \mathsf{O}(k^{m_2-1}|\lambda_2|^k)$, elementwise, where m_2 is the algebraic multiplicity of λ_2 and $\mathbf{Q}_\infty$ is the matrix whose rows are equal to the limiting distribution [25, Theorem 1.2]. Thus the convergence is geometric with rate $|\lambda_2|$. For $\mathbf{PAM_{250}}$, we numerically compute $|\lambda_2| = 0.53$. However, finding the eigenvalues of $\mathbf{P}$, other than 1, amounts to finding the roots of a polynomial of degree 19 analytically. Since there is no algebraic way to find the roots of such a polynomial, the following inequality, due to Deutsch and Zenger, gives an upper bound for λ_2 [41]:

$$|\lambda_2| \leq \tfrac{1}{2} \max_{i,j} \left\{ p_{i,i} + p_{j,j} - p_{i,j} - p_{j,i} + \sum_{\substack{k \\ k\neq i,j}} |p_{i,k} - p_{j,k}| \right\}. \tag{19}$$

Applying (19) to the probability transition matrix $\mathbf{P}$, in Figure 2, leads to $|\lambda_2| \leq 1 - \frac{1}{2}\alpha$. □

PROOF OF THEOREM 1. Denote by $\min^+ I$ the minimum of the strictly positive elements of the set I. Theorem 1 follows from Theorem 4.10 of [25], which states that if the sequence $\mathbf{T}_{p,k}$ is scrambling, for all $k \geq 1$, and $\min^+_{i,j} q(k)_{i,j} \geq \gamma > 0$ uniformly for all $k \geq 1$, then weak ergodicity obtains at a uniform geometric rate for all $p \geq 1$. Let

$$\gamma = \min_{1\leq k\leq N} \left\{ \min^+_{i,j} \mathbf{PAM}(k)_{i,j} \right\}.$$

Then we have $\min^+_{i,j} \mathbf{PAM}(k)_{i,j} \geq \gamma > 0$ uniformly for all $k \geq 1$. Observe that the main assumption in Theorem 1 is the finite number of PAM matrices. From the proof of Theorem 4.10 of [25], it follows that the convergence rate is geometric with parameter $(1 - \gamma^t)^{1/t}$. □

PROOF OF THEOREM 2. From the probability transition matrix $\mathbf{P}(k)$, depicted in Figure 2, we have

$$\min_{i,j}^{+} p_{i,j}(k) = \min\{1 - 9\alpha(k), \tfrac{1}{6}\alpha(k)\}. \tag{20}$$

From the boundedness of the mutation rate $\alpha(k)$ $(0 < a \leq \alpha(k) \leq b < 1)$, we obtain

$$\min_{i,j}^{+} p_{i,j}(k) \geq \min\left\{\frac{a}{6}, 1 - 9b\right\} = \gamma, \tag{21}$$

uniformly on k. Let $\mathbf{e}_k$ be the unique stationary distribution of $\mathbf{P}(k)$. We have $\mathbf{e}_k = \mathbf{s_1}$ in (5), for all $k \geq 1$. In particular, the sequence of vectors $\{\mathbf{e}_k\}_{k\geq 1}$ converges to $\mathbf{s_1}$. Since $\mathbf{T}_{p,k}$ have no zero column, the strong ergodicity property follows from Theorem 4.15 of [25]. The rate of convergence follows from Theorem 4.10 of [25]. □

PROOF OF PROPOSITION 3. Write $T = M + S$, where S is the total number of nucleotide introns in the gene. Then, assuming a total of $n \geq 1$ mutations in the gene, the probability of error P_e is given by

$$P_e(S) = \sum_{k=1}^{n} \binom{n}{k} \frac{M^k S^{n-k}}{(M+S)^n} = 1 - \left(\frac{S}{M+S}\right)^n. \tag{22}$$

The derivative of P_e with respect to the variable S is

$$P_e^{'}(S) = -\frac{nMS^{n-1}}{(M+S)^{n+1}} < 0, \qquad \text{for all} \quad n \geq 1. \tag{23}$$

Hence P_e is a decreasing function of the intron length for all $n \geq 1$. Moreover, (22) is independent of the intron–exon structure in the gene. □

PROOF OF PROPOSITION 4. Let x_k denote the start position of the kth exon in the genome. We have

$$P_e = 1 - \prod_{k=1}^{K} Pr(\text{“0 error in exon } k\text{”}), \tag{24}$$

where

$$\begin{aligned} Pr(\text{“0 error in exon } k\text{”}) &= \sum_{n=0}^{T-l_k} e^{-\lambda l_k} \Pr(\text{“}n \text{ errors outside } l_k) \\ &= \sum_{n=0}^{T-l_k} e^{-\lambda l_k} \left\{ \sum_{i=0}^{n} Pr(\text{“i errors} \in [1, x_k - 1]\text{”}) \right. \\ &\qquad \left. \times Pr(\text{“}(n-i) \text{ errors} \in [x_k + l_k, T]\text{”}) \right\} \\ &= \sum_{n=0}^{T-l_k} e^{-\lambda l_k} \left(\sum_{i=0}^{n} e^{-\lambda(x_k-1)} \frac{(\lambda(x_k-1))^i}{i!} e^{-\lambda(T-x_k-l_k+1)} \right. \\ &\qquad \left. \times \frac{(\lambda(T-x_k-l_k+1))^i}{(n-i)!} \right) \end{aligned} \tag{25}$$

$$= \sum_{n=0}^{T-l_k} e^{-\lambda T} \sum_{i=0}^{n} \frac{\lambda^n}{n!} \binom{n}{i} (x_k - 1)^i (T - x_k - l_k + 1)^{n-i}$$

$$= e^{-\lambda T} \sum_{n=0}^{T-l_k} \frac{\lambda^n}{n!} (T - l_k)^n.$$

From (25) and (24), we obtain

$$P_e = 1 - e^{-\lambda K T} \prod_{k=1}^{K} \sum_{n=0}^{T-l_k} \frac{\lambda^n (T - l_k)^n}{n!}. \tag{26}$$

□

PROOF OF PROPOSITION 5. Let f_X be a unimodal density which reaches its mode at x_0. Then $f_X(x - x_0)$ reaches its mode at 0. We have

$$x_0^* = \operatorname*{argmax}_{x_0} \int_{M/K-\rho}^{M/K+\rho} f_X(x - x_0)\, dx. \tag{27}$$

By continuity of f_X, we have

$$\left|(x - x_0) - \left(\frac{M}{K} - x_0\right)\right| < \rho \quad \Rightarrow \quad \left|f_X(x - x_0) - f_X\left(\frac{M}{K} - x_0\right)\right| < \varepsilon, \tag{28}$$

for some $\varepsilon > 0$. So,

$$\left|x - \frac{M}{K}\right| < \rho \quad \Rightarrow \quad f_X\left(\frac{M}{K} - x_0\right) - \varepsilon < f_X(x - x_0) < f_X\left(\frac{M}{K} - x_0\right) + \varepsilon. \tag{29}$$

So,

$$\operatorname*{argmax}_{x_0} 2\rho \left(f_X\left(\frac{M}{K} - x_0\right) - \varepsilon\right) \le x_0^* \le \operatorname*{argmax}_{x_0} 2\rho \left(f_X\left(\frac{M}{K} - x_0\right) + \varepsilon\right).$$

Since $f_X(x - x_0)$ reaches its mode at 0, we obtain $x_0^* = M/K$. □

PROOF OF PROPOSITION 6.

$$P_e = 1 - \prod_{k=1}^{K} Pr(\text{“0 error in exon } k\text{”})$$

$$= 1 - \prod_{k=1}^{K} \int_0^\infty Pr(\text{“0 error in exon } k \mid \text{its length is } l\text{”})\, p(l)\, dl$$

$$= 1 - \prod_{k=1}^{K} \int_0^\infty e^{-\lambda l} p(l)\, dl$$

$$= 1 - \left(\int_0^\infty e^{-\lambda l} p(l)\, dl\right)^K.$$

□

References

[1] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, Wiley-Interscience, New York, 1991.

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

[3] L. L. Gatlin, *Information Theory and the Living System*, Columbia University Press, New York, 1972.

[4] H. P. Yockey, *Information Theory and Molecular Biology*, Cambridge University Press, Cambridge, 1992.

[5] R. R. Roldan, P. B. Galvan, and J. L. Oliver, Application of information theory to DNA sequence analysis: a review, *Pattern Recognition*, vol. 29, no. 7, pp. 1187–1194, 1996.

[6] E. E. May, Analysis of Coding Theory Based Models for Initiating Protein Translation in Prokaryotic Organisms, Ph.D. thesis, North Carolina State University, Raleigh, NC, March 2002.

[7] H. P. Yockey, *Information Theory*, *Evolution and the Origin of Life*: *Fundamentals of Life*, Elsevier, New York, 2002.

[8] J. M. G. Rosen, Investigation of coding structure in DNA, in *Proceedings of the IEEE International Conference on Acoustics*, *Speech*, *and Signal Processing*, Hong Kong, April 2003, pp. 361–364.

[9] E. E. May, M. A. Vouk, and D. L. Bitzer, Classification of escherichia coli k-12 ribosome binding sites, *IEEE Engineering in Medicine and Biology Magazine*, vol. 25, no. 1, pp. 90–97, January 2006.

[10] G. Battail, Does information theory explain biological evolution, *Europhysics Letters*, vol. 40, no. 3, pp. 343–348, 1997.

[11] D. R Forsdyke, Are introns in-series error-detecting sequences?, *Journal of Theoretical Biology*, vol. 93, pp. 861–866, 1981.

[12] G. Battail, Should genetics get an information-theoretic education?, *IEEE Engineering in Medicine and Biology Magazine*, vol. 25, no. 1, pp. 34–45, January 2006.

[13] R. D. Wood, M. Mitchell, J. Sgouros, and T. Lindahl, Human DNA repair genes, *Science*, vol. 291, no. 5507, pp. 1284–1289, 2001.

[14] L. S. Liebovitch, Y. Tao, A. T. Todorov, and L. Levine, Is there an error correcting code in the base sequence in DNA?, *Biophysical Journal*, vol. 71, no. 3, pp. 1539–1544, 1996.

[15] G Battail, Replication decoding revisited, in *Proceedings of the IEEE Information Theory Workshop*, ENST, Paris, April 2003, pp. 1–5.

[16] W. Gilbert, S.J. De Souza, and M. Long, Origin of genes, *Proceedings of the National Academy of Sciences*, vol. 94, pp. 7698–7703, July 1994.

[17] Mouse Genome Sequencing Consortium, Initial sequencing and comparative analysis of the mouse genome, *Nature*, vol. 420, pp. 520–562, December 2002.

[18] R. D. Wood, M. Mitchell, J. Sgouros, and T. Lindahl, Human DNA repair genes, *Science*, vol. 291, no. 5507, pp. 1284–1289, 2001.

[19] J. M. Berg, J. L. Tymoczko, and L. Stryer, *Biochemistry*, Michelle Julet, 2002.

[20] M. Dayhoff, R. Schwartz, , and B. Orcutt, A model of evolutionary change in proteins, in M. Dayhoff, ed., *Atlas of Protein Sequence and Structure*, vol. 5, pp. 345–352, National Biomedical Research Foundation, Silver Spring, MD, 1978.

[21] D. T. Jones, W. R. Taylor, and J. M. Thornton, The rapid generation of mutation data matrices from protein sequences, *Bioinformatics*, vol. 8, pp. 275–282, 1992.

[22] T. H. Jukes, R. Holmquist, and H. Moise, Amino acid composition of proteins: selection against the genetic code, *Science*, vol. 189, pp. 50–51, 1975.

[23] L. J. Gross, B. C. Mullin, and S. E. Riechert, Alternative routes to quantitative literacy for the life sciences, July 2000. See www.tiem.utk.edu/∼gross/bioed.

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

[25] E. Seneta, *Non-Negative Matrices and Markov Chains*, Springer-Verlag, Berlin, 1981.

[26] S. Ohno, *Evolution by Gene Duplication*, Springer-Verlag, Berlin, 1970.

[27] W. H. Li and D. Grauer, *Fundamentals of Molecular Biology*, Sinauer, Sunderland, MA, 1994.

[28] J. M. Comeron and M. Kreitman, Negative correlation between intron length and recombination rate, *Genetics*, vol. 156, no. 3, pp. 1175–1190, 2000.

[29] A. Papoulis and S. U. Pillai, *Probability, Random Variables and Stochastic Processes*, McGraw-Hill, New York, 1991.
[30] C. Burge, Identification of Genes in Human Genomic, Ph.D. thesis, Stanford University, March 1997.
[31] W. Gilbert, Gene structure and evolutionary theory, in *New Perspective on Evolution*, 1991, pp. 155–163.
[32] Z. Zhang and M. Gerstein, Patterns of nucleotide substitution, insertion and deletion in the human genome inferred from pseudogenes, *Nucleic Acids Research*, vol. 31, no. 18, pp. 5338–5348, 2003.
[33] C. K. Peng, S. V. Buldyrev, A. L. Goldberger, S. Havlin, F. Sciortino, M. Simons, and H. E. Stanley, Long-range correlations in nucleotide sequences, *Nature*, vol. 356, pp. 168–170, March 1992.
[34] S. V. Buldyrev, A. L. Goldberger, S. Havlin, R. N. Mantegna, M. E. Matsa, C. K. Peng, M. Simons, and H. E. Stanley, Long-range correlation properties of coding and noncoding DNA sequences: genbank analysis, *Physical Review E*, vol. 51, no. 5, pp. 5084–5091, May 1995.
[35] C. Madalena, A. L Goldberger, and C. K. Peng, Multiscale entropy analysis of biological signals, *Physical Review E*, vol. 71, 2005.
[36] B. V. Gnedenko and A. N. Kolmogorov, *Limit Distributions for Sums of Independent Random Variables*, Addison-Wesley, Reading, MA, 1954.
[37] V. M. Zolotarev, *One-Dimensional Stable Distributions*. Volume 65 of Translations of Mathematical Monographs, American Mathematical Society, Providence, RI, 1986.
[38] J. P. Nolan, Parametrizations and modes of stable distributions, *Statistics and Probability Letters*, vol. 38, pp. 187–195, 1998.
[39] S. Henikoff and J. G. Henikoff, Amino acid substitution matrices from protein blocks, *Proceedings of the National Academy of Sciences of the USA*, 1992, vol. 89, pp. 10915–10919.
[40] G. Frobenius, Uber matrizen aus nicht negativen elementen, *Königlich Preussische Akademie der Wissenschaften* (*Berlin*), pp. 456–477, 1908.
[41] E. Deutsch and C. Zenger, Inclusion domains for the eigenvalues of stochastic matrices, *Numerische Mathematik*, vol. 18, pp. 182–192, 1971.