Proceedings of the 28th IEEE
EMBS Annual International Conference
New York City, USA, Aug 30-Sept 3, 2006

Biological Evolution: Distribution and Convergence Analysis of Amino Acids

Nidhal Bouaynaya and Dan Schonfeld

Abstract— **We propose a protein communication system where the transmitted messages are protein sequences and the encoded message is the DNA. A series connection of the protein communication channel is equivalent to a channel through time: the channel of evolution. We study the evolutionary dynamics of this channel in both cases of constant and time-varying point mutation rate. We establish, using matrix analysis, that stochastic messages sent through the channel of evolution are received according to a fixed probability distribution, which is independent of the original message.**

I. Introduction

An essential ingredient to any life form is the existence of an information storage and processing system within it. Furthermore, life can be understood and described as a communication process through time. Thus, the development of a mathematical model to capture the genetic information storage and transmission apparatus, during cell division, is important for many research areas such as intron research, aging theories and evolutionary studies.

The standard in the mathematical biology community today is to model the biological information transfer as a communication system, where the input is the DNA sequence and the output is the amino acid chain in the protein [1], [2], [3]. Even though the DNA-Protein system faithfully reproduces the biological flow of information, it fails to explain the different elements in a proper biological communication system. Specifically:

- The DNA-Protein system is 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). It is, therefore, incorrect to view the translation between DNA sequences and proteins as a communication system. The DNA-protein system is a transformation between the 4-letter alphabet message in the DNA and the 20-letter alphabet message in the amino acid polypeptide. The genetic code dictates this transformation. Thus, from a communication point of view, the DNA-Protein system corresponds to the decoding system;
- The DNA-protein system views the DNA as the message source and hence completely neglects the true nature of the DNA sequence as the encoded information, which is well established in molecular biology, even though there is no encoding process in biology;
- In the DNA-Protein system, the source DNA generates the genome according to a specific stochastic process, which uses a 4-letter alphabet. Hence, the DNA-Protein system cannot explain the current structure of DNA, e.g. presence of non-coding DNA and the size of the genetic alphabet.

The closest notion, found in the literature, to a mathematical abstraction of a protein communication channel has been presented by May et al. [4] who introduce a communication model with a virtual genetic encoder, where the DNA is the encoded information and the proteins are the decoded information; yet they somehow fail to model the information source as a source of amino acid alphabets. In this paper, we model the transmission of information, during cell replication or asexual reproduction,as a protein communication system with a single source generating the protein set of the parent. It is important to emphasize however that, in this view, we are not supporting the theory of a biological protein-protein genetic code. The proposed protein communication system is a mathematical model of information transmission during cell division. This model does not support either the theories of proteins-first or nucleotides-first at the origin of life. It 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. In fact, the proposed biological 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 $\rightarrow$ MPEG $\rightarrow$ MPEG $\rightarrow$ video" even though the process "video $\rightarrow$ MPEG" never takes place. Biological organisms have resolved the real communication problem, i.e. "protein $\rightarrow$ protein", by ensuring that organisms maintain both proteins and DNA. Therefore, the "protein $\rightarrow$ 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 [5].

Analysis of 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. In particular, it requires the use of multi-user information theory and dis-

1-4244-0033-3/06/$20.00 ©2006 IEEE.

tributed coding and will not be discussed in this paper.

The protein communication system is shown in Fig. 1(a). This system structure suggests a strong isomorphism with engineering communication systems. However, there are two main differences between the genetic information processing system and the communication engineer's 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's system, the biological communication system is not designed to minimize transmission errors. In the absence of errors, evolution will not be possible. Fig. 1(b) summarizes the analogy between an engineering communication system for video transmission and the protein communication system.

The protein communication channel is time-dependent: thermal noise, radioactivity and cosmic rays are sources of errors and they occur with a probability that is a function of time regardless of the number of replications of the DNA. A series connection of the protein communication channel is equivalent to a channel through time: "the channel of evolution". In this paper, we will investigate the behavior of this channel. Specifically, we will address the following questions: (1) Given an infinitely small probability of error at each generation of cell replication, how are the cell offsprings related to their ancestral mother cell after a large number of generations? (2) Given an initial distribution of amino acids, how does this distribution evolve with time? Is there an equilibrium distribution? If yes, what is the rate of convergence to this equilibrium distribution and what are the biological implications of such equilibrium?

II. Protein Communication Channel

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

$$\text{Pr } (P_j|P_i) = \prod_{k=1}^{N} \text{Pr } (a_k^j|a_k^i), \tag{1}$$

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

In this paper, we use two different probability transition matrices: PAM_{250} probability transition matrix [6] [1] and a first-order Markov transition probability matrix, $\mathbf{P}$. $\mathbf{P}$ 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 64 codons are equally probable and from Baye's rule, we obtain the following formula for the probability of a transition from amino acid a to amino acid $\hat{a}$,

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

where $\{c_1, \cdots, c_n\}$, (resp. $\{b_1, \cdots, b_m\}$), 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 1 point mutations, we retain only the terms of the first degree in $\alpha(k)$. The probability transition matrix $\mathbf{P}$ is displayed in Fig. 2. The amino acids are alphabetically ordered by their one-letter standard abbreviations, e.g., $p_{1,1} = \text{Pr}(A|A)$.

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 is given by

$$\mathbf{p}_k = \mathbf{p}_0\mathbf{Q}(1)\mathbf{Q}(2)\cdots\mathbf{Q}(k), \tag{2}$$

where $\mathbf{Q} \in \{\text{PAM}_{250}, \mathbf{P}\}$. Observe that $\mathbf{P}$ takes into account all possible mutations between amino acids whether they are accepted or rejected by natural selection whereas the PAM transition matrix is estimated from phylogenetic trees of protein sequences and hence takes into account the accepted mutations only.

III. Constant Point Mutation Rate

In this section, we assume that the point mutation rate is constant over time, i.e., $\alpha(k) = \alpha$, for all $k \geq 0$. Hence, Eq. (2) becomes

$$\mathbf{p}_k = \mathbf{p}_0\mathbf{Q}^k. \tag{3}$$

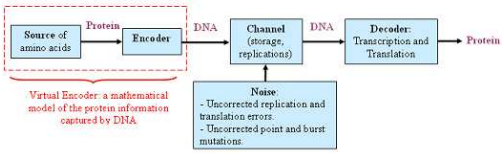
Proposition 1: Consider an initial probability distribution of the amino acids at time 0, $\mathbf{p}_0$ (some amino acids might have an initial zero probability of occurrence). Then, the probability distribution of the amino acids converges, over time, towards a stationary distribution given by $\mathbf{s_1}$ if $\mathbf{Q} = \mathbf{P}$ and $\mathbf{s_2}$ if $\mathbf{Q} = \text{PAM}_{250}$, where

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

$$\mathbf{s_2} = (0.087,\ 0.041,\ 0.042,\ 0.048,\ 0.034,\ 0.039,\ 0.051,\ 0.091,\ 0.033,\ 0.036,\ 0.083,\ 0.08,\ 0.014,\ 0.038,\ 0.053,\ 0.07,\ 0.06,\ 0.0089,\ 0.028,\ 0.064).$$

Proof: The probability transition matrices $\mathbf{P}$ and PAM_{250} are irreducible and aperiodic. Therefore, from the Perron-Frobenius theorem [7], 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. $\{\mathbf{p}_0\ \text{PAM}_{250}^k\}_{k\in\mathbb{N}}$) 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 solution of the linear system $\mathbf{s_1P} = \mathbf{s_1}$ (resp. $\mathbf{s_2}\ \text{PAM}_{250} = \mathbf{s_2}$), subject to $\mathbf{s_1 1} = 1$ (resp. $\mathbf{s_2 1} = 1$), where $\mathbf{1}$ is the column vector with all its entries equal to 1. ■

[1] The PAM_{250} probability transition matrix is the PAM mutation probability matrix for the evolutionary distance of 250 PAMs and should not be confused with the PAM Log odds matrix corresponding to the same evolutionary distance. The PAM_{250} transition probability matrix is shown in [6, Fig. 83]



(a)

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.

(b)

Fig. 1. (a) Protein communication system; (b) Comparison between the engineering communication system and the protein communication system.

$$\begin{pmatrix}
1-6\alpha & 0 & \frac{\alpha}{2} & \frac{\alpha}{2} & 0 & \alpha & 0 & 0 & 0 & 0 & 0 & 0 & \alpha & 0 & 0 & \alpha & \alpha & \alpha & 0 & 0 & 0 \\
0 & 1-8\alpha & 0 & 0 & \alpha & \alpha & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \alpha & 2\alpha & 0 & 0 & \alpha & \alpha & \alpha \\
\alpha & 0 & 1-8\alpha & 2\alpha & 0 & \alpha & 0 & 0 & 0 & 0 & 0 & \alpha & 0 & 0 & 0 & 0 & 0 & \alpha & 0 & \alpha & 0 \\
\alpha & 0 & 2\alpha & 1-8\alpha & 0 & \alpha & 0 & 0 & \alpha & 0 & 0 & 0 & 0 & \alpha & 0 & 0 & 0 & \alpha & 0 & 0 & \alpha \\
0 & \alpha & 0 & 0 & 1-8\alpha & 0 & 0 & \alpha & 0 & 3\alpha & 0 & 0 & 0 & 0 & 0 & \alpha & 0 & \alpha & 0 & \alpha & 0 \\
\alpha & \frac{\alpha}{2} & \frac{\alpha}{2} & \frac{\alpha}{2} & 0 & 1-6\alpha & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{3\alpha}{2} & \frac{\alpha}{2} & 0 & \alpha & \frac{\alpha}{4} & 0 & \frac{\alpha}{4} \\
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 & 0 & \frac{2\alpha}{3} & 0 & 0 & 1-7\alpha & \frac{\alpha}{3} & \frac{4\alpha}{3} & \alpha & \frac{2\alpha}{3} & 0 & 0 & \frac{\alpha}{3} & \frac{2\alpha}{3} & \alpha & \alpha & 0 & 0 & 0 \\
0 & 0 & 0 & \alpha & 0 & 0 & 0 & \frac{\alpha}{2} & 1-8\alpha & 0 & \frac{\alpha}{2} & 2\alpha & 0 & \alpha & \alpha & 0 & \alpha & 0 & 0 & 0 & \alpha \\
0 & 0 & 0 & 0 & \alpha & 0 & \frac{\alpha}{3} & \frac{2\alpha}{3} & 0 & 1-6\alpha & \frac{\alpha}{3} & 0 & \frac{2\alpha}{3} & \frac{\alpha}{3} & \frac{2\alpha}{3} & \frac{\alpha}{3} & 0 & \alpha & \frac{\alpha}{6} & 0 & \frac{\alpha}{2} \\
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 & 0 & \alpha & 0 & 0 & 0 & \alpha & \alpha & 2\alpha & 0 & 0 & 1-8\alpha & 0 & 0 & 0 & \alpha & \alpha & 0 & 0 & \alpha & 0 \\
\alpha & 0 & 0 & 0 & 0 & 0 & \frac{\alpha}{2} & 0 & 0 & \alpha & 0 & 0 & 1-6\alpha & \frac{\alpha}{2} & \alpha & \alpha & \alpha & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & \alpha & 0 & 0 & 2\alpha & 0 & \alpha & \alpha & 0 & 0 & \alpha & 1-8\alpha & \alpha & 0 & 0 & 0 & 0 & 0 & \alpha \\
0 & \frac{\alpha}{3} & 0 & 0 & 0 & \alpha & \frac{\alpha}{3} & \frac{\alpha}{6} & \frac{\alpha}{3} & \frac{2\alpha}{3} & \frac{\alpha}{6} & 0 & \frac{2\alpha}{3} & \frac{\alpha}{3} & 1-6\alpha & \alpha & \frac{\alpha}{3} & 0 & \frac{\alpha}{3} & 0 & \frac{\alpha}{3} \\
\frac{2\alpha}{3} & \frac{2\alpha}{3} & 0 & 0 & \frac{\alpha}{3} & \frac{\alpha}{3} & 0 & \frac{\alpha}{3} & 0 & \frac{\alpha}{3} & 0 & \frac{\alpha}{3} & \frac{2\alpha}{3} & 0 & \alpha & 1-\frac{20\alpha}{3} & \alpha & 0 & \frac{\alpha}{6} & \frac{\alpha}{3} & \frac{\alpha}{2} \\
\alpha & 0 & 0 & 0 & 0 & 0 & 0 & \frac{3\alpha}{4} & \frac{\alpha}{2} & 0 & \frac{\alpha}{4} & \frac{\alpha}{2} & \alpha & 0 & \frac{\alpha}{2} & \frac{3\alpha}{2} & 1-6\alpha & 0 & 0 & 0 & 0 \\
\alpha & 0 & \frac{\alpha}{2} & \frac{\alpha}{2} & \frac{\alpha}{2} & \alpha & 0 & \frac{3\alpha}{4} & 0 & \frac{3\alpha}{2} & \frac{\alpha}{4} & 0 & 0 & 0 & 0 & 0 & 0 & 1-6\alpha & 0 & 0 & 0 \\
0 & 2\alpha & 0 & 0 & 0 & \alpha & 0 & 0 & 0 & \alpha & 0 & 0 & 0 & 0 & 2\alpha & \alpha & 0 & 0 & 1-9\alpha & 0 & 2\alpha \\
0 & \alpha & \alpha & 0 & \alpha & 0 & \alpha & 0 & 0 & 0 & 0 & \alpha & 0 & 0 & 0 & \alpha & 0 & 0 & 0 & 1-8\alpha & 2\alpha \\
0 & \frac{2\alpha}{3} & 0 & \frac{2\alpha}{3} & 0 & \frac{\alpha}{3} & 0 & 0 & \frac{2\alpha}{3} & \alpha & 0 & 0 & 0 & \frac{2\alpha}{3} & \frac{2\alpha}{3} & \alpha & 0 & 0 & \frac{2\alpha}{3} & \frac{4\alpha}{3} & 1-\frac{23\alpha}{3}
\end{pmatrix}$$

Fig. 2. $\mathbf{P}$: a first-order Markov probability transition matrix between amino acids. Only the terms of the first degree in $\alpha(k)$ are retained.

Observe that $\mathbf{s_1}$ is proportional to the number of codon assignment in proteins. Jukes et al. [8] studied 68 representative proteins from eukaryotic, prokaryotic and viruses. They computed the following distribution vector, $\mathbf{r}$ for the 20 amino acids,

$$\mathbf{r} = (\tfrac{5.3}{61}, \tfrac{1.3}{61}, \tfrac{3.6}{61}, \tfrac{3.3}{61}, \tfrac{2.3}{61}, \tfrac{4.8}{61}, \tfrac{1.4}{61}, \tfrac{3.1}{61}, \tfrac{4.1}{61}, \tfrac{4.7}{61}, \tfrac{1.1}{61}, \tfrac{3}{61}, \tfrac{2.5}{61}, \tfrac{2.4}{61}, \tfrac{2.6}{61}, \tfrac{4.5}{61}, \tfrac{3.7}{61}, \tfrac{4.2}{61}, \tfrac{0.8}{61}, \tfrac{2.3}{61}). \quad (4)$$

Since PAM_{250} estimates the rate of accepted mutations only, $\mathbf{s_2}$ is closer, on average, to $\mathbf{r}$ than $\mathbf{s_1}$. 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. In both cases, the limiting distribution of the amino acids are not uniform. Hence, some amino acids will be more abundant than others and consequently, evolution will have a higher probability of generating certain organisms. 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. The mean experimental and limiting distributions, for each class, are very close except for the class of amino acids corresponding to 6 codons obtained from the limiting distribution using the probability transition matrix $\mathbf{P}$. The reason is that Arginine, which is coded by 6 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 [2].

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: $\{\mathbf{p}_0\mathbf{Q}^k\}_{k\geq 1}$ converges at a geometric rate with parameter $|\lambda_2|$, where $\begin{cases} |\lambda_2| = 0.53, & \text{if } \mathbf{Q} = \text{PAM}_{250}; \\ |\lambda_2| \leq 1 - \frac{1}{2}\alpha, & \text{if } \mathbf{Q} = \mathbf{P}. \end{cases}$
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, as a consequence, proves that no evolution is possible if $\alpha = 0$.

Proof: The matrix $\mathbf{Q} \in \{\mathbf{P}, \text{PAM}_{250}\}$ 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 [9, Theorem 1.2]. Thus the convergence is geometric with rate $|\lambda_2|$. For PAM_{250}, we numerically compute $|\lambda_2| = 0.53$. However, Finding the eigenvalues of $\mathbf{P}$, other than 1, amounts to analytically finding the roots of a polynomial of degree 19. Since there is no algebraic way to find the roots of such a polynomial, the following inequality, due to Deutsch & Zenger, gives an upper bound for λ_2 [10]:

$$|\lambda_2| \leq \frac{1}{2} \max_{i,j} \{p_{i,i} + p_{j,j} - p_{i,j} - p_{j,i} + \sum_{\substack{k \\ k \neq i,j}} |p_{i,k} - p_{j,k}|\}. \quad (5)$$

Applying Eq. (5) to the probability transition matrix $\mathbf{P}$, in Fig 2, leads to $|\lambda_2| \leq 1 - \frac{1}{2}\alpha$. ■

IV. 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): [9] The forward products $\mathbf{T}_{p,k}$ are said to be *weakly ergodic* if $t_{i,s}^{p,k} - t_{j,s}^{p,k} \xrightarrow{k\to\infty} 0$ for each i, j, s, p. 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 obtains, then the limit row vector $\mathbf{v}_p = \{v_j^{(p)}\}$ is a probability vector and is independent of $p \geq 0$, i.e., $\mathbf{v}_p = \mathbf{v}$ [9]. 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: [9] 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 since every transition matrix at time k, $\mathbf{Q}(k)$, is scrambling, then so is $\mathbf{T}_{p,k}, p \geq 0$.

Theorem 1: Consider a finite number of PAM matrices denoted by PAM(1),$\cdots$, PAM(N), where PAM(i) can be PAM_1 or PAM_{160} or PAM_{250}, etc, for all $i = 1, \cdots, N$. Consider the sequence: $\mathbf{T}_{p,k} = \mathbf{t}_{p+1}\mathbf{t}_{p+2}\cdots\mathbf{t}_{p+k}$, where each $\mathbf{t}_i \in \{\text{PAM}(1), \cdots \text{PAM}(N)\}$. 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 equal). Then, $\mathbf{T}_{p,k}$ is weakly ergodic at a uniform geometric rate for all $p \geq 0$. So the sequence $\{\mathbf{p}_k\}_{k\geq 1}$, in Eq. (2), tends to a sequence of distributions independently of $\mathbf{p_0}$.

Proof: Denote by $\min^+ I$ the minimum of the strictly positive elements of the set I. Theorem 1 follows from [9, Theorem 4.10], 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

$$\gamma = \min_{1\leq k\leq N} \{ \overset{+}{\min_{i,j}} \ \text{PAM}(k)_{i,j}\}.$$

Then we have $\min_{i,j}^+ \text{PAM}(k)_{i,j} \geq \gamma > 0$ 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 [9, Theorem 4.10], it follows that the convergence rate is geometric with parameter $(1-\gamma^t)^{\frac{1}{t}}$. ∎

If we approximate the matrices PAM_k by PAM_1^k, the sequence $\mathbf{T}_{p,k} = \text{PAM}^{p+1}\text{PAM}^{p+2}\cdots\text{PAM}^{p+k}$ becomes strongly ergodic. In particular, the sequence $\{\mathbf{p}_k\}_{k\geq 1}$, in Eq. (2), converges to the limiting distribution $\mathbf{s_2}$.

Theorem 2: Consider a point mutation rate, $\alpha(k)$, which is bounded uniformly on k, i.e., $0 < a \leq \alpha(k) \leq b < 1$. 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 Eq. (2), converges towards the stationary distribution $\mathbf{s_1}$ independently of the initial distribution $\mathbf{p}_0$. Moreover, the convergence rate is at least geometric with parameter $(1-\gamma^t)^{\frac{1}{t}}$, where $\gamma = \min\{\frac{a}{6}, 1-9b\}$.

Proof: From the probability transition matrix $\mathbf{P}(k)$, depicted in Fig. 2, we have $\min_{i,j}^+ p_{i,j}(k) = \min\{1-9\alpha(k), \frac{1}{6}\alpha(k)\}$. From the boundedness of the mutation rate $\alpha(k)$, we obtain $\min_{i,j}^+ p_{i,j}(k) \geq \min\{\frac{a}{6}, 1-9b\} = \gamma$, uniformly on k. Let $\mathbf{e}_k$ be the unique stationary distribution of $\mathbf{P}(k)$. We have, $\mathbf{e}_k = \mathbf{s_1}$ 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 [9, Theorem 4.15]. The rate of convergence follows from [9, Theorem 4.10]. ∎

The time-varying point mutation rate analysis implies, in particular, that the original transmitted message is somehow lost through the channel of evolution; this is an information theoretic proof of the darwinian theory since the human protein set, for instance, is the received message of a primitive bacteria protein set transmitted through the channel of evolution at the beginning of life.

V. Conclusion

We can obtain similar results with the BLOSUM [11] probability transition matrix constructed from the log-odds BLOSUM matrix. The convergence of the probability transition matrix shows that a parent organism will be unrelated to its offsprings after infinitely many generations no matter how small the initial point mutation rate is as long as it is non-zero. The rate of convergence quantifies the speed of this divergence. The limiting distribution $\mathbf{s_1}$ shows that, if all mutations were accepted, the asymptotic abundance of amino acids in nature would be proportional to their codon assignment. The discrepancy between this limiting distribution and the natural abundance can be related to the relative survival of the amino acids after they mutate.

References

[1] L. Gatlin, *Information Theory and the Living Systems*. Cambridge University Press, 1972.

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

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

[4] M. V. E. May, D. Bitzer, and D. Rosnick, "A coding theory framework for genetic sequence analysis," in *Workshop on genomic Signal processing and Statistics (GENSIPS)*, 2002, p. 11.

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

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

[7] G. Frobenius, "Uber matrizen aus nicht negativen elementen," *S.B. Preuss. Akad. Wiss.*, pp. 456–477, 1912.

[8] 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.

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

[10] E. Deutsch and C. Zenger, "Inclusion domains for the eigenvalues of stochastic matrices," *Numerische Math.*, vol. 18, pp. 182–192, 1971.

[11] S. Henikoff and J. G. Henikoff, "Amino acid substitution matrices from protein blocks," in *Proc. Natl. Acad. Sci. USA.*, vol. 89, 1992, pp. 10915 – 10919.