# ANALYSIS OF PROTEIN EVOLUTION AS A COMMUNICATION SYSTEM

*Nidhal Bouaynaya and Dan Schonfeld*

Electrical and Computer Engineering Department, University of Illinois at Chicago.

## ABSTRACT

We propose a protein communication system where the transmitted messages are protein sequences and the encoded message is the DNA. We study the evolutionary dynamics of this channel in both cases of constant and time-varying point mutation rate. We prove that the distribution of amino acids converges, at a geometric rate, to a limiting distribution.

## 1. INTRODUCTION

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 [1]. The protein message is encoded into the DNA sequence before transmission through the channel. The encoding process 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. DNA storage and replication is part of the biological communication medium, or physical channel, which introduces errors to the communication system. The decoding process, called *translation* in biology , is accomplished based on the well-known genetic code. The translation process can make errors. To simplify the communication model, these errors are incorporated as part of the physical channel.

## 2. PROTEIN COMMUNICATION CHANNEL

Assuming a memoryless channel, it can be easily shown that the protein communication channel is characterized by the probability transition matrix, $\mathbf{Q}(k) = \{q_{i,j}(k)\}_{1\leq i,j\leq 20}$, at time $k$, of the amino acids.

[1]In this paper, by abuse o f notation, we denote by 'protein' or 'protein sequence' the polypeptide chain of amino acids which forms the 3-D folded protein.

In this paper, we use two different probability transition matrices: PAM and a first-order Markov transition probability matrix, $\mathbf{P}$. We construct $\mathbf{P}$ 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 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 $\hat{a}$,

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

where $\{c_1,\cdots,c_n\}$, resp. $\{b_1,\cdots,b_m\}$, are the codons of the received ($\hat{a}$), resp. transmitted ($a$), amino acid and $h(b_j,c_i)$ is the hamming distance between codon $b_j$ and codon $c_i$. Since burst mutations are less likely to happen than 1 point mutation and for computational efficiency, we retain only the terms of the first degree in $\alpha(k)$.

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

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

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

## 3. 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 $k \geq 0$. Equation (1) becomes

$$\mathbf{p}_k = \mathbf{p}_0\mathbf{Q}^k. \tag{2}$$

**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 [2] converges, over time,*

[2]The amino acids are ordered alphabetically by their one-letter standard abbreviations. For instance, $\lim_{k\rightarrow\infty}$ Pr (L)= $\lim_{k\rightarrow\infty}$ Pr (R)=$\lim_{k\rightarrow\infty}$ Pr (S)=$\frac{6}{61}$.

*towards a stationary distribution given by* $\mathbf{s_1}$ *if* $\mathbf{Q} = \mathbf{P}$ *and* $\mathbf{s_2}$ *if* $\mathbf{Q} = PAM_{250}$*, where*

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

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

The proof of Proposition 1 follows from the Perron-Frobenius theorem. Jukes et al. [1] computed the following distribution, $\mathbf{r}$, from a study of representative proteins from eukaryotic, prokaryotic and viruses;

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

Since $PAM_{250}$ estimates the rate of accepted mutations, $\mathbf{s_2}$ is closer, on average, to $\mathbf{r}$ than $\mathbf{s_1}$.

**Proposition 2** $\{\mathbf{p}_0\mathbf{Q}^k\}_{k\geq 1}$ *converges at a geometric rate with parameter* $|\lambda_2|$*, where* $\begin{cases} |\lambda_2| = 0.53, & \text{if } \mathbf{Q} = PAM_{250}; \\ |\lambda_2| \leq 1 - \frac{1}{2}\alpha, & \text{if } \mathbf{Q} = \mathbf{P}. \end{cases}$

As a consequence, no evolution is possible if $\alpha = 0$.

**proof 1** *Let the eigenvalues of* $\mathbf{Q}$ *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* $\lambda_2$ *and* $\mathbf{Q}_\infty$ *is the matrix whose rows are equal to the limiting distribution [2, Theorem 1.2]. The following inequality gives an upper bound for* $\lambda_2$ *of the probability transition matrix* $\mathbf{P}$ *[3]:*

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

## 4. 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** *[2] 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 obtains and the* $t_{i,s}^{p,k}$ *themselves tend to a limit for all* $i, s, p$*, then we say* strong ergodicity *obtains.*

**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* $\mathbf{t}_i \in \{PAM(1), \cdots PAM(N)\}$*. 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,* $T_{p,k}$ *is weakly ergodic at a uniform geometric rate for all* $p \geq 0$*. So the sequence* $\{\mathbf{p}_k\}_{k\geq 1}$*, in Eq. (1), tends to a sequence of distributions independently of* $\mathbf{p_0}$*.*

**proof 2** *Denote by* $\min^+$ *the minimum of the strictly positive elements. Let* $\gamma = min_{1\leq k\leq N}(min_{i,j}^+ PAM(k)_{i,j})$*. Then we have* $\min^+{}_{i,j}\, PAM(k)_{i,j} \geq \gamma$ *uniformly for all* $k \geq 1$*. Then, theorem 1 follows from [2, Theorem 4.10]. The convergence rate is geometric with parameter* $(1-\gamma^t)^{\frac{1}{t}}$*.*

If we approximate the matrices $\text{PAM}_k$ by $\text{PAM}_1^k$, 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. (1), 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. So, the sequence* $\{\mathbf{p}_k\}_{k\geq 1}$*, in Eq. (1), 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 3** *We have* $\min^+{}_{i,j}\, p_{i,j}(k) = \min\{1-9\alpha(k), \frac{1}{6}\alpha(k)\}$*. Hence,* $\min^+{}_{i,j}\, p_{i,j}(k) \geq \gamma$*, 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 then from [2, Theorem 4.15]. The rate of convergence follows from [2, Theorem 4.10].*

## 5. CONCLUSION

We can obtain similar results with the BLOSUM 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 is related to the relative survival of the amino acids after they mutate.

## 6. REFERENCES


[1] 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.

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

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