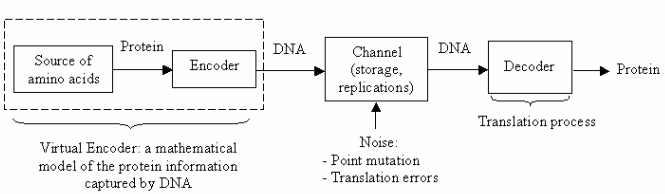
Protein Communication System: Evolution and Genomic Structure

Nidhal Bouaynaya and Dan Schonfeld
University of Illinois at Chicago · nbouay1@uic.edu · dans@uic.edu

Protein Communication Channel

Adaptive evolution has fashioned living organisms as agents of information acquisition, analysis, storage, and transmission. How has this happened? How have living systems evolved to handle the same problems with which we are confronted in this so-called Information Age: problems of information storage and processing, problems of transmission and reliability ?

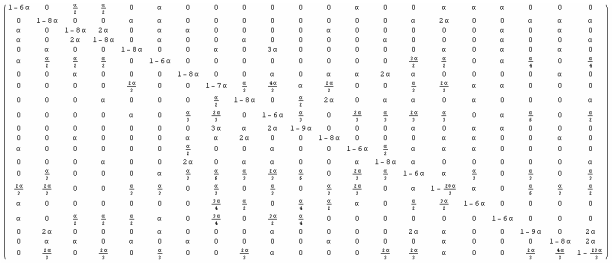
Protein Communication Channel



Analogy and Differences with a Communication Engineering System

Protein Communication System	Communication Engineering System
Set of proteins of the cell	Video
DNA	MPEG
No biological encoder	Encoder
Translation process	Decoder
? (The genomic structure provides the answer)	Objective function: minimize the probability of error

The protein communication channel is uniquely characterized by the probability transition matrix, $Q(k) = \{q_{i,j}(k)\}$, $1 \le i,j \le 20$, at time k of the amino acids.



P: a first-order Markov probability transition matrix between amino acids. Only the terms of the first degree in $\mathfrak{b}(k)$ are retained. For display clarity, the dependence on the time k has been omitted.

$$p_k = p_0\, Q(1)\, Q(2) \cdots Q(k),$$

where $\mathbf{Q} \in \{\mathrm{PAM}, \mathbf{P}\}$.

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.

Evolution: Constant Point Mutation Rate

Proposition 1: *(Convergence of the amino acid distribution)* Consider an initial probability distribution of the amino acids at time 0, p_0 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 $\mathbf{Q} = \mathrm{PAM}_{250}$, where

$\mathbf{s}_1 = (0.062,\ 0.0312,\ 0.0312,\ 0.0312,\ 0.062,\ 0.0312,\ 0.046,\ 0.0312,\ 0.093,\ 0.0156,\ 0.0312,\ 0.0625,\ 0.0312,\ 0.0937,\ 0.0937,\ 0.0625,\ 0.0625,\ 0.0156,\ 0.0312)$

$\mathbf{s}_2 = (0.087,\ 0.041,\ 0.042,\ 0.048,\ 0.034,\ 0.039,\ 0.051,\ 0.091,\ 0.033,\ 0.036,\ 0.083,\ 0.08,\ 0.014,\ 0.038,\ 0.053,\ 0.07,\ 0.06,\ 0.0089,\ 0.028,\ 0.064)$.

The experimental distribution is

$\mathbf{r} = (0.0868, 0.0213, 0.059, 0.054, 0.0377, 0.0786, 0.023, 0.0508, 0.0672, 0.077,\ 0.018,\ 0.0491, 0.041, 0.0393, 0.0426, 0.0737, 0.0606, 0.0688, 0.0131, 0.0377)$.

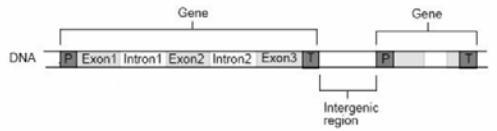
Proposition 2: *(Rate of Convergence)* $\{p_0 Q_k\}_{k \ge 1}$ converges at a **geometric rate** with parameter $\mathfrak{m}_2$, where $\begin{cases} \mathfrak{m}_2 = 0.53, & \text{if } \mathbf{Q} = \mathrm{PAM}_{250}; \\ \mathfrak{m}_2 \nearrow 1 - \mathfrak{b}/2, & \text{if } \mathbf{Q} = \mathbf{P}. \end{cases}$

Evolution: Time-Varying Point Mutation Rate

Theorem 1:*(Weak Ergodicity result)* Consider a finite number of PAM matrices denoted by PAM(*1*),· · · , 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 $\mathbf{ti} \in \{\mathrm{PAM}(1), \cdots \mathrm{PAM}(N)\}$. That is at each time k, the probability transition matrix is some PAM matrix. Then, $\mathbf{T}_{p,k}$ is weakly ergodic at a uniform geometric rate for all $p \neq 0$. So the sequence $\{\mathbf{p}_k\}_{k \ge 1}$tends to a sequence of distributions independently of p_0.

Theorem 2: *(Strong Ergodicity Result)* Consider a point mutation rate, $\mathfrak{b}(k)$, which is bounded uniformly on k, i.e., $0 < a \nearrow \mathfrak{b}(k) \nearrow b < 1$. Then the products $\mathbf{T}_{p,k} = \mathbf{P}_{p+1} \cdots \mathbf{P}_{p+k}$ are strongly ergodic. Thus, the sequence $\{p_k\}_{k \ge 1}$converges towards the stationary distribution $\mathbf{s}_1$ independently of the initial distribution p_0. Moreover, the convergence rate is at least geometric.

Genomic Structure: Proposed Theory



Coding and Non-Coding Regions in DNA

We show that introns protect coding regions in the DNA sequence from frequent errors in the way 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.

Genomic Structure: Deterministic Analysis

Under the assumption of a Poisson ($\mathfrak{m}$) noise, we obtain the probability of error:

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

Taking the derivative of P_e e with respect to l_k, we obtain the following coupled system for the optimal exon lengths:

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

An obvious solution is obtained when $\mathbf{l_k = M / K}$ **for all** $k = 1, \cdots, K$.

The asymmetric distribution, which best approximates $\delta_{M/K}$ 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!

Genomic Structure:Stochastic Analysis

Probability of Error Analysis Let *p(l)* be the continuous distribution of the length of exons.

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

Stochastic Optimization Problem:

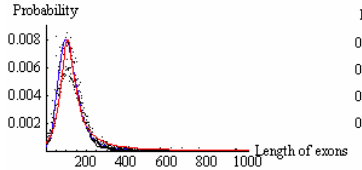
$$p^*(l) = \arg\max_{p(l)} \int_0^\infty e^{-\lambda l} p(l)\, dl \quad \text{subject to} \quad \begin{cases} i) \int_0^\infty p(l)\, dl \\ ii) \int_0^\infty l^{1+\alpha} p(l)\, dl \ge l_0 \end{cases}$$

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

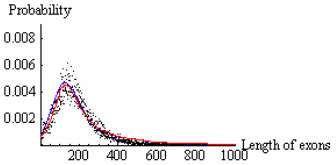
Random Walk Model $X_N = \sum_{i=1}^{N} l_i, \quad l_i \sim f(l) = \alpha l^{-(1+\alpha)}$

$$\lim_{x \to \infty} p(x \mid \alpha, \beta, \gamma, \delta) = \frac{C}{x^{1+\alpha}}$$

Experimental Results



Homo-Sapiens (Human)



Apis Mellifera (Honey bee)