INFORMATION-THEORETIC BOUNDS OF EVOLUTIONARY PROCESSES MODELED AS A PROTEIN COMMUNICATION SYSTEM

Liuling Gong, Nidhal Bouaynaya and Dan Schonfeld*

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

ABSTRACT

In this paper, we investigate the information theoretic bounds of the channel of evolution introduced in [1]. The channel of evolution is modeled as the iteration of protein communication channels over time, where the transmitted messages are protein sequences and the encoded message is the DNA. We compute the capacity and the rate-distortion functions of the protein communication system for the three domains of life: Achaea, Prokaryotes and Eukaryotes. We analyze the tradeoff between the transmission rate and the distortion in noisy protein communication channels. As expected, comparison of the optimal transmission rate with the channel capacity indicates that the biological fidelity does not reach the Shannon optimal distortion. However, the relationship between the channel capacity and rate distortion achieved for different biological domains provides tremendous insight into the dynamics of the evolutionary processes. We rely on these results to provide a model of protein sequence evolution based on the two major evolutionary processes: mutations and unequal crossover.

Index Terms— Biological communication system; Channel capacity; Rate-distortion theory.

1. INTRODUCTION

The genetic information storage and transmission apparatus resembles engineering communication systems in many ways: The genomic information is digitally encoded in the DNA. By decoding genes into proteins, organisms come into being. The protein communication system, proposed in [1], [2] and shown in Fig. 1, is a communication model of the genetic information storage and transmission apparatus. The protein communication system abstracts a cell as a set of proteins and models the process of cell division as an information communication system between protein sets. Using this mathematical model of protein communication, the problem of a species' evolution will be represented as the iteration of a communication channel over time.

The genome is viewed as the joint source-channel encoded message of the protein communication system and hence can be investigated in the context of engineering communication codes. In particular, it is legitimate to ask at what rate can the genomic information be transmitted. And what is the average distortion between the transmitted message and the received message at this rate? Shannon's channel capacity theorem states that, by properly encoding the source, a communication system can transmit information at a rate that is as close to the channel capacity as one desires with an arbitrarily small transmission error. Conversely, it is not possible to reliably transmit at a rate greater than the channel capacity. The theorem, however, is not constructive and does not provide any help in designing such codes. In the case of biological communication systems, however, evolution has already designed the code for us. The encoded message is the DNA sequence. Comparison of the genomic transmission rate with the channel capacity will reveal whether the genomic code is efficient from an information theoretic perspective. However, even if the channel capacity is not exceeded, we are assured that biological communication systems do not rely on codes that produce negligible errors since the level of distortion presented must account for evolutionary processes. It is, therefore, interesting to ask ourselves whether biological communication systems maintain an optimal balance between the transmission rate and the desired distortion level needed to support adaptive evolution. Rate-distortion theory analyzes the optimal tradeoff between the transmission rate, $R(D)$, and distortion, D, in noisy communication channels. Given the fidelity, D, present in biological communication systems, comparison of the genomic transmission rate with the optimal rate $R(D)$ can be used to determine whether or not the genomic code achieves the optimal rate-distortion criteria. Moreover, by equating the optimal rate $R(D)$ with the channel capacity, C, we can determine whether the biological fidelity, D, reaches the Shannon optimum distortion. In this paper, we will only compare the channel capacity and rate distortion functions of a single source memoryless protein communication system, modelling asexual reproduction. The two-source protein communication system, modelling sexual reproduction, is more involved mathematically and will not be addressed here.

*Nidhal Bouaynaya is currently in the Department of Systems Engineering at the University of Arkansas at Little Rock.

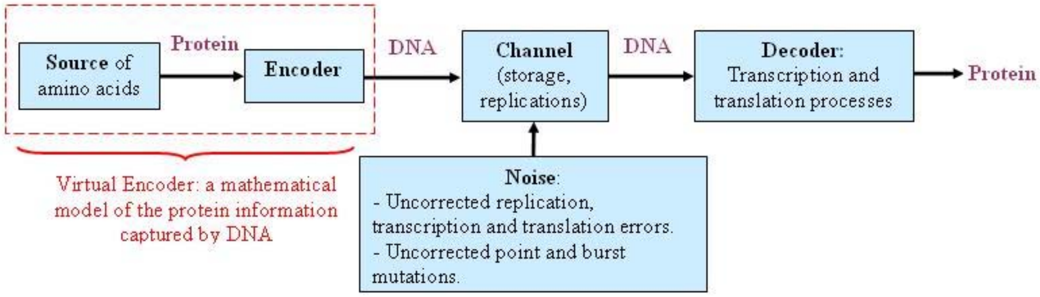

Source of amino acids
Protein
Encoder
DNA
Channel (storage, replications)
DNA
Decoder: Transcription and translation processes
Protein
Noise:
- Uncorrected replication, transcription and translation errors.
- Uncorrected point and burst mutations.
Virtual Encoder: a mathematical model of the protein information captured by DNA

Fig. 1. Protein communication system

2. PROTEIN CHANNEL CAPACITY

Assuming a first-order Markov channel, the protein communication channel is characterized by the probability transition matrix, $Q = \{q_{i,j}\}_{1\leq i,j\leq 20}$, of the amino acids. In this paper, we use two different probability transition matrices: Dayhoff's Point Accepted Mutation (PAM) matrices [3], and a first-order Markov transition probability matrix **P** [1], [2]. An element of a PAM matrix, M_{ij}, gives the probability that the amino acid in row i will be replaced by the amino acid in column j after a given evolutionary interval which is interpreted as the evolutionary distance of PAM matrices. The first-order Markov transition probability matrix **P** is constructed from the genetic code using a point mutation rate α, which represents the probability of a base interchange of any one nucleotide and is assumed to be constant over time (see [1] for the computation of **P**).

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

$$C = \max_{p\in P^n} I(p,Q) = \max_{p\in P^n} \sum_j \sum_k p_j Q_{jk} \log \frac{Q_{jk}}{\sum_k p_j Q_{jk}}, \tag{1}$$

where $P^n = \{p \in \mathbb{R}^n : p_j \geq 0 \ \forall j; \sum_j p_j = 1\}$ is the set of all probability distributions on the channel input, Q is the probability transition matrix of the channel, and $I(p,Q)$ is known as the mutual information between the channel input and output. Evaluation of the channel capacity involves solution of a convex programming problem. In most cases, analytic solutions cannot be found. Blahut [4] suggested an iterative algorithm for computing the channel capacity.

Figure 2(a) (resp. 2(b)) shows the capacity of the protein communication system as a function of the evolutionary distance of PAM matrices (resp. point mutation rate α). As expected, the channel capacity decreases to zero as the evolutionary distance or the point mutation rate α increases. This result has different ramifications on bioinformatics than on communication engineering: In engineering, it is interpreted as a loss of information after a great number of transmissions. And no information can be obtained at the output after infinite transmissions. The reason is that, in communications, only the initial message is used to convey information and not the channel. In bioinformatics, on the other hand, the output message captures the information of the channel (i.e. the mutations) regardless of the initial message. In particular, a parent organism cannot transmit reliably (channel capacity zero) its genetic information to its offspring of many generations no matter how small the point mutation rate is as long as it is not zero. After a long enough time of evolution, the final distribution of amino acids in offspring only depends on the channel characteristics regardless of the parent organism. It is also interesting to observe that organisms with lower mutation rates have higher channel capacity, and therefore their genetic information can be reliably transmitted at a higher rate.

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

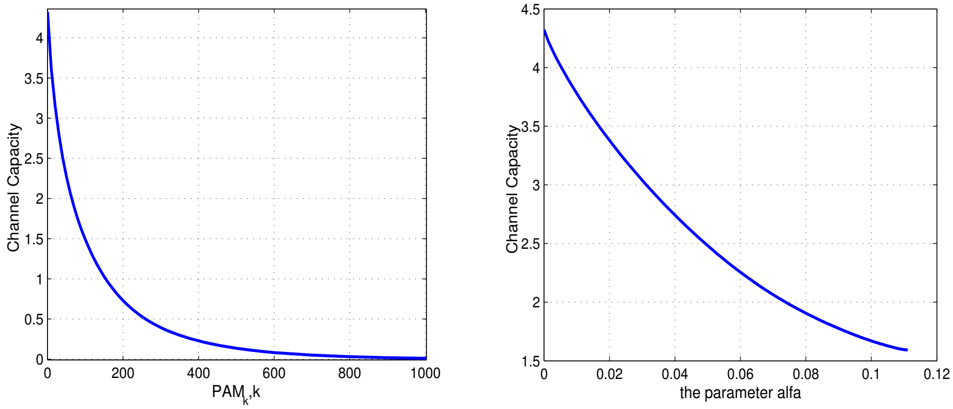
3. PROTEIN RATE DISTORTION

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

$$R(D) = \min_{Q\in Q_D} I(p,Q) = \min_{Q\in Q_D} \sum_j \sum_k p_j Q_{jk} \log \frac{Q_{jk}}{\sum_k p_j Q_{jk}}, \tag{2}$$

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

We define the distortion between a pair of amino acids as their distance in the Principal Component Analysis (PCA) plane obtained from 7 physico-chemical properties (volume,



(a) (b)

Fig. 2. Channel Capacity: (a) Channel capacity v.s. the evolutionary distance of PAM matrices; (b) Channel capacity v.s. the point mutation rate α

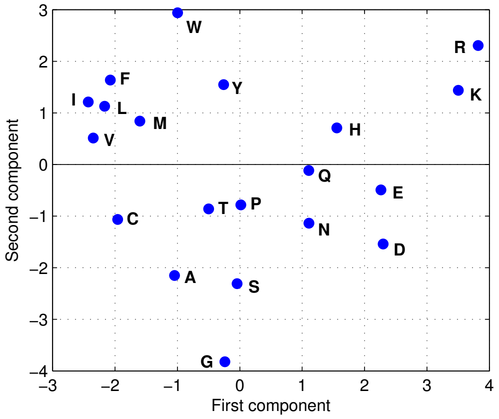


Fig. 3. Plot of the amino acids on the first two components of the PCA analysis. The amino acids are labelled by their one-letter standard abbreviations.

bulkiness, polarity, PH index, hydrophobicity and surface area). The amino acid data was obtained from [Chapter 2] [5]. The result of PCA analysis is shown in Fig 3. The amino acid probability distributions in Archaea, Bacteria and Eukaryote were experimentally computed in [6]. Using Blahut's algorithm for rate-distortion functions [4], we compute the rate-distortion curves for Archaea, Bacteria and Eukaryote. They are displayed in Fig. 4.

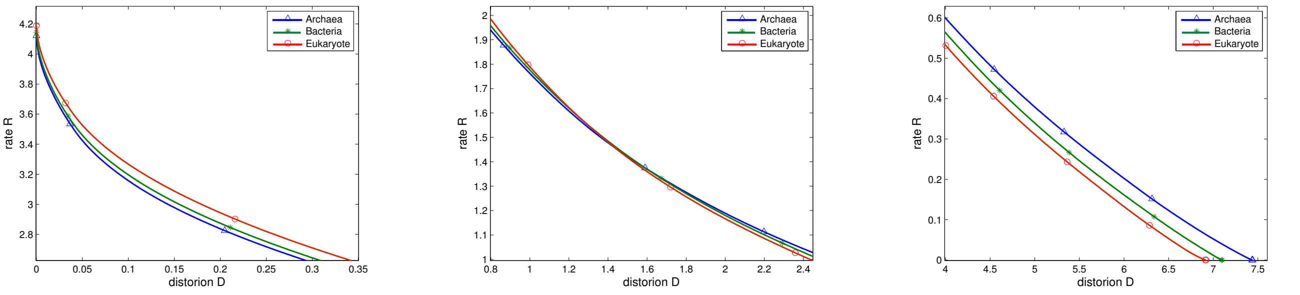
Figure 4 reveals two distinct regions: a low distortion region ($0 \leq D \leq 1.4$) and a high distortion region ($1.4 \leq D \leq 7.5$). In the low-distortion region, the R-D curve of Eukaryotes is the highest followed by Bacteria, then Archaea, i.e., we have

$$R(D)_{Ar} < R(D)_{Ba} < R(D)_{Eu}, \quad \forall\, 0 < D < 1.4, \qquad (3)$$

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

$$R(D)_{Eu} < R(D)_{Ba} < R(D)_{Ar}, \quad \forall\, 1.4 < D < 7.5. \qquad (4)$$

The distortion can be associated with the evolutionary distance. That is a low distortion region would correspond to small evolutionary distances, whereas the high distortion region corresponds to larger evolutionary distances. It is quite interesting to observe that for small evolutionary distances (or at the beginning of life), Archaea was the most efficient organism from an information theoretic perspective, followed by Bacteria then Eukaryotes. Specifically, given a fixed transmission rate (of the genetic information), Archaea would have the least distortion. At about $D \approx 1.4$, the three R-D curves intersect and reverse orders. So, for longer evolutionary distances, Eukaryotes maintain the most biological fidelity among



(a) (b) (c)

Fig. 4. Rate-distortion curves for Archaea, Bacteria and Eukaryotes: (a) low distortion region; (b) intersection region; (c) high distortion region.

the three domains.

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

$$D = \sum_j \sum_k p_j q_{jk} \rho_{jk}, \tag{5}$$

where $Q = \{q_{i,j}\}$ is the probability transition matrix of the channel, $p = \{p_j\}$ is the distribution of the channel input and $\rho_{i,j}$ is the distortion between amino acids i and j. By trial and error Dayhoff et al. [3] found that the 250 PAM matrix works well for scoring of actual protein sequences. At this evolutionary distance (250 substitutions per hundred residues) only one amino acid in five remains unchanged. Table 1 displays the actual average distortion for Archaea, Bacteria and Eukaryotes, where PAM$_{250}$ was used as the probability transition matrix of the channel. Observe that the biological rate-distortion values $R(D)$, corresponding to the average distortions given in Table 1, are less than the Shannon channel capacity ($C = 0.8197 > R(D)$). So, from rate-distortion theory, we can ascertain that the genetic information is encoded so that the system reproduces the initial input with fidelity D. In particular, the biological communication system does not rely on codes that produce negligible errors since the level of distortion present must account for evolutionary processes.

The formula of the rate-distortion function given in Eq. (2) is valid only for discrete-time stationary and memoryless sources. For discrete-time stationary sources with memory, Wyner and Ziv derived bounds for their rate-distortion function [7] as follows:

$$R(D) - \Delta \leq R^*(D) \leq R(D) \ , \tag{6}$$

where $R(D)$ is the rate-distortion function of the memoryless source with the same marginal statistics, Δ is a measure of the memory of the source and is independent of the distortion measure and the distortion value D. So, the R-D curves for a source with memory are always shifted down compared to the R-D curves of the corresponding memoryless source. Moreover, the shift is a function of the source and not the distortion. Thus, the biological rate-distortion values $R^*(D)$ corresponding to the average distortions given in Table 1 are still less than the Shannon channel capacity and the bounds on the R-D curves still exhibit the same reversal phenomenon depicted in Fig. 4.

Table 1. Average Rate-Distortion for three domains of life

	Archaea	Bacteria	Eukaryote
Distortion	9.1491	8.9964	8.8979

Table 2. Scaled norm of odd moments

	Archaea	Bacteria	Eukaryote
scaled norm	5.0653	73.5401	1.0000

3.1. Evolutionary Model: Amino Acid Distribution

It is well known from information theory that the Gaussian input maximizes the mutual information in an additive Gaussian noise [8], i.e.,

$$I(X; X + Z^*) \leq I(X^*; X^* + Z^*), \tag{7}$$

where $I(a, b)$ is the mutual information between input a and output b, X is the input, Z is the channel noise and * denotes Gaussianity. We will show that the amino acid distribution in Eukaryotes is "more Gaussian" than Bacteria and Archaea. Since a distribution is uniquely characterized by the set of its moments and given that the odd moments of the Gaussian distribution are identically zero, we compute the odd moments of the amino acid distribution for the three branches of life [6]. Table 2 displays the scaled norm of the first 4 odd moments (3rd, 5th, 7th and 9th) for Archaea, Bacteria and Eukaryotes. The odd moments norm of Archaea (resp. Bacteria) are 5 (resp. 73) times higher than Eukaryotes, asserting that the amino acid distribution of Eukaryotes is "more Gaussian" than the two other groups of life. To explain the low-distortion region in Fig.4 and the switching-over of the R-D curves, we have to dig deeper into the evolutionary processes, which shaped the three groups of life.

3.2. Evolutionary Process: Mutation and Crossover

It is widely accepted today that the main driving forces of evolution are mutations and unequal crossover [1]. Furthermore, Archaea and Bacteria rely mostly on mutations for adaptability and survival. So, we can fairly postulate that mutations drive the evolution of Archaea and Bacteria whereas unequal crossovers drive the evolution of Eukaryotes. A mutation involves one nucleotide or a very short sequence of nucleotides. Therefore, it induces much less modifications to the genome sequence than any unequal crossover. So at the beginning of evolution, the distortion caused by mutations is small compared to the distortion caused by unequal crossovers. However, with time, mutations accumulate much faster than the rare unequal crossovers. So, the distortion caused by mutations exceeds, over time, the distortion caused by unequal crossovers. This implies higher fidelity, over time, in Eukaryotes than Bacteria and Archaea. For example, assume that mutations and unequal crossovers follow a Non Homogeneous Poisson Process (NHPP) within the genome. The NHPP process is a Poisson Process with a time-dependent rate parameter, $\lambda(t)$. The Probability that there are n events in the interval $(r, r+s)$ is calculated as follows

$$P(N(r,r+s)-N(r)=n)=\frac{e^{-\int_r^{r+s}\lambda(t)\,dt}(\int_r^{r+s}\lambda(t))^n}{n!} \tag{8}$$

It can be shown that there exist parameters $\lambda(t)_{mutation}$ and $\lambda(t)_{unequal\ crossover}$ representing different functions of time which induce the trend of the R-D curves observed in Fig. 4.

4. CONCLUSION

By modeling evolution as the iteration of a protein communication system over time, we were able to study it from an information theoretic perspective. Investigation of the biological communication channel capacity and the rate-distortion curves of the three branches of life, Archaea, Bacteria and Eukaryotes, reveals that the biological fidelity D does not reach the Shannon optimum distortion. Furthermore, we relied on these results to provide an evolutionary model of the three groups of life based on mutations and unequal crossovers.

5. REFERENCES

[1] N. Bouaynaya and D. Schonfeld, "Protein communication system: Evolution and genomic structure," *Algorithmica, Special issue on Algorithmic Methodologies for Processing Protein Structures, Sequences and Networks*, to appear.

[2] N. Bouaynaya and D. Schonfeld, "Biological evolution: Distribution and convergence analysis of amino acids," in *28th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBS'06)*, August 2006, pp. 2045–2048.

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

[4] R. Blahut, "Computation of channel capacity and rate-distortion functions," *IEEE Transactions on Information Theory*, vol. 18, pp. 460–472, July 1972.

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

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

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

[8] T. M. Cover and J. A. Thomas, *Element of Information Theory*, Wiley-Interscience, New York, 2nd edition, 2006.

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