**BMC
Bioinformatics**

The empirical codon mutation matrix as a communication channel

Dawit Nigatu, Attiya Mahmood and Werner Henkel[*]

Abstract

Background: A number of evolutionary models have been widely used for sequence alignment, phylogenetic tree reconstruction, and database searches. These models focus on how sets of independent substitutions between amino acids or codons derive one protein sequence from its ancestral sequence during evolution. In this paper, we regard the Empirical Codon Mutation (ECM) Matrix as a communication channel and compute the corresponding channel capacity.

Results: The channel capacity of 4.1875 bit, which is needed to preserve the information determined by the amino acid distribution, is obtained with an exponential factor of 0.26 applied to the ECM matrix. Additionally, we have obtained the optimum capacity achieving codon distribution. Compared to the biological distribution, there is an obvious difference, however, the distribution among synonymous codons is preserved. More importantly, the results show that the biological codon distribution allows for a "transmission" at a rate very close to the capacity.

Conclusion: We computed an exponential factor for the ECM matrix that would still allow for preserving the genetic information given the redundancy that is present in the codon-to-amino acid mapping. This gives an insight how such a mutation matrix relates to the preservation of a species in an information-theoretic sense.

Background

Markov models for the protein sequence evolution have been widely used for the past 40 years. These evolutionary matrices highlight the most common mutational changes between amino acids and codons. Protein sequence evolution has been investigated at both amino acid and codon levels. The evolutionary matrices on the basis of amino acids are widely used for sequence alignments and phylogenetic tree construction. As more than one codon encode to the same amino acid, it is easy to estimate alignments in amino acids as compared to codons.

Codon-level models demonstrate the mutational changes among the codons. This gives us more information by highlighting the tendency of mutations between codons encoding the same amino acid (synonymous changes) as well as the mutational effects between codons that code for different amino acids (non-synonymous changes). As codons are the smallest genetic information unit in protein-encoding regions, it is obvious to model mutations by a codon-based communications channel model highlighting all codon-to-codon changes in nature.

Substitution matrices define the rate at which one amino acid in the protein sequence is changed into another amino acid. Dayhoff et al. [1] estimated the first such model in 1972, resulting in the widely used point accepted mutations (PAM) matrix. It is computed by counting the mutations in the closely related proteins. Henikoff and Henikoff proposed the block substitution matrix (BLOSUM) for divergent protein sequences, which uses log-likelihood ratios to construct scoring matrices from the transition matrices between amino acids [2]. Later on, Whelan and Goldman (WAG) proposed a novel approach to estimate amino acid replacement matrices from a large database of aligned protein sequences in 2001 [3]. It combines the estimation of transition and scoring matrices by a maximum-likelihood approach that accounts for the phylogenies of sequences within each training alignment.

As the codon (a tri-nucleotide) is the basic genetic information that directly encodes the amino acid as the building block of proteins, we have used the first empirical codon substitution matrix (ECM) in our analysis. This was proposed by Schneider et al., where sequences of five

*Correspondence: w.henkel@jacobs-university.de
School of Engineering and Science, Jacobs University Bremen, Campus Ring 1,
D-28759 Bremen, Germany

vertebrates were aligned and the number of codon substitutions were counted among them [4]. According to conversations with the authors, it is estimated that these mutations on average happened in roughly 300 Million years.

Yockey was one of the first to model and describe a central dogma using information theoretic tools [5]. He viewed the flow of information from DNA or RNA to proteins as a communication system and employed entropy, rate, and capacity calculations with a transition matrix he developed by considering base changes of equal probability. A detailed analysis of the application of information theory to molecular biology can be found in his book [6]. Relatively recently, L. Gong, N. Bouaynaya, and D. Schonfeld have proposed a communication model for protein evolution [7]. They used the amino acid based PAM matrix and a matrix they produced, similar to Yockey's, as a communication channel and performed capacity calculations over it.

We computed an exponential factor for the ECM matrix that would still allow for preserving the genetic information given the redundancy that is present in the codon-to-amino acid mapping. This gives an insight on how such a mutation matrix relates to the preservation of a species in an information-theoretic sense.

For the underlying capacity computation, we used the Arimoto-Blahut algorithm [8,9] to determine the input distribution that maximizes the mutual information.

Methods

In order to compute the mutation probability in the ECM matrix, 17502 alignments of sequences from five vertebrate genomes yielded 8.3 million aligned codons from which the number of substitutions between codons were counted. This matrix has 64×64 entries stating the mutation probability of each codon to every other codon. Basically, the substitution from sense codons to stop codons is not included in the ECM matrix, which makes the matrix block diagonal with a 61×61 matrix for coding codons and a 3×3 entries for substitutions between stop codons. Therefore, we will consider only substitutions between coding codons and regard the ECM matrix as 61×61. From the communication perspective, this mutation matrix describes channel transition probabilities $\mathbf{P}(y|x)$.

There is also another matrix in [4], which gives the actual count of substitutions observed. From this substitution count matrix $\mathbf{C}$, we obtained the biological probability distribution of the codons as

$$\mathbf{p}_x = \frac{\sum\limits_{j} C_{ij}}{\sum\limits_{i}\sum\limits_{j} C_{ij}} \, . \tag{1}$$

Thereafter, we combined the codons which encode for the same amino acid and computed the probability distribution of amino acids, denoted $\mathbf{p}_a$. Using this distribution, the to be preserved information content of the 64 codons representing the 20 amino acids can be computed as

$$R_{20} = -\sum_{i=1}^{20} \mathbf{p}_a(i) \log_2(\mathbf{p}_a(i)). \tag{2}$$

According to Shannon's channel coding theorem, a communication through a noisy channel of capacity C at an information rate of R is possible with an arbitrarily small probability of error, if $R < C$ [10]. Hence, the channel capacity has to, at least, exceed the value of R_{20}.

In communication systems, the channel capacity is determined by maximizing the mutual information between input (X) and output (Y) over the input probability distribution $\mathbf{p}_x$.

$$C = \sup_{\mathbf{p}_x} I(X;Y). \tag{3}$$

$I(X;Y)$ is the mutual information which measures the mutual dependence between input and output distributions, and is defined as

$$I(X;Y) = H(Y) - H(Y|X), \tag{4}$$

where $H(Y)$ is the entropy of the codon distribution at the output of the ECM "channel", and $H(Y|X)$ is the conditional entropy, called prevarication or irrelevance.

However, in the system we are considering, the input distribution (i.e., probability distribution of codons) is not something to adjust. It is defined by nature. Therefore, we determine the channel capacity corresponding to the mutation "channel" matrix for a biological codon frequency obtained by Eq. (1). $H(Y)$ is computed as

$$H(Y) = -\sum_{i=1}^{61} p_{y_i} \log_2(p_{y_i}), \tag{5}$$

where p_{y_i} is the output probability distribution of the ith codon. The conditional entropy $H(Y|X)$ between input and output distribution of codons is computed as

$$H(Y|X) = -\sum_{i=1}^{61} p(x_i) \sum_{j=1}^{61} p(y_j|x_i) \log_2 p(y_j|x_i). \tag{6}$$

$p(y_j|x_i)$ is the conditional probability between codons, which is given by the empirical codon mutation (ECM) matrix.

We now compute, what exponent of the ECM matrix would be needed to make the capacity just match the required rate obtained by Eq. (2). Hereto, we use the singular value decomposition (SVD) yielding

$$\left[\mathbf{P}(y|x)\right]^F = \mathbf{U}(\Sigma)^F \mathbf{V}^*, \tag{7}$$

where $\mathbf{U}, \mathbf{V}$ are unitary matrices, Σ is a diagonal matrix with nonnegative real numbers in the diagonal, and F is an exponent to be fine-tuned. The value of the exponent is changed in steps from zero to one. A value of 1 means the original ECM matrix is used.

Moreover, we would like to find the optimum codon distribution by solving Eq. (3) and compare it with the biological distribution. For solving the optimization problem, the Arimoto-Blahut algorithm was employed [8,9]. The Arimoto-Blahut algorithm is an iterative numerical algorithm that monotonically converges to the capacity value. To compute the capacity, it is starting from any arbitrary input probability distribution $\mathbf{p}_x$ (usually uniform) and performs the following two steps until the algorithm converges.

1. Compute a quantity related to the mutual information per input symbol

$$c(x_j) := \exp \sum_k p(y_k|x_j) \log \frac{p(y_k|x_j)}{\sum_j p(x_j)p(y_k|x_j)}. \quad (8)$$

This results from a Lagrange multiplier step in [9].
2. Update the input probability distribution according to

$$p(x_j) = \frac{p(x_j)c(x_j)}{\sum_j p(x_j)c(x_j)} . \quad (9)$$

The termination criteria is based on the lower and upper bounds of the channel capacity,

$$\log \left(\sum_j p(x_j)c(x_j) \right) \leq C \leq \log \left(\max_{x_j} c(x_j) \right). \quad (10)$$

The iterations are terminated when the upper and lower bounds are equal up to a certain accuracy.

Once the optimized codon distribution is obtained using the Arimoto-Blahut algorithm, to note the similarity with the biological distribution, we applied the so called Kullback-Leibler divergence (D_{KL}) [11]. D_{KL} is a quantitative measure of how similar a probability distribution P is to a model distribution Q, and is defined as

$$D_{KL}(P||Q) = \sum_i P_i \log_2 \frac{P_i}{Q_i}. \quad (11)$$

D_{KL} is non-negative and gives a zero result when the distributions are perfectly matched. Technically speaking, D_{KL} measures the average number of extra bits required (coding penalty) for using a code based on Q instead of P.

Results and discussion

The to be preserved information content of the amino acids, using the amino acid distribution and computed according to Eq. (2) is 4.1875 bit, which is less than the maximum value of $\log_2(20) = 4.3219$ bit. Likewise, the required rate obtained by using the amino acid probability distribution provided by King & Jukes in [12], derived from 5492 residues of 53 vertebrate polypeptides is 4.2033 bit. Thus, it is reasonable to look for a capacity that is at least greater than 4.1875. Hence, using the biological codon distribution in the five vertebrates obtained by using Eq. (1), we stepwise reduced the exponent of the ECM matrix until it satisfies the rate requirement. Furthermore, we used the Arimoto-Blahut algorithm to find the optimal input probability distribution of the 61 codons to maximize the mutual information and compare it with the biological distribution of codons. The optimal capacity-achieving codon distribution and the observed biological codon distribution are both shown in Figure 1. The corresponding values are also tabulated in Table 1 and Table 2.

The capacity obtained by optimizing the codon distribution, the mutual information based on the observed biological codon distribution, and the required rate are shown together in Figure 2. When the exponent of the ECM matrix is reduced, the output codon distribution changes and the prevarication $H(Y|X)$ will be smaller. As a result, the capacity increases. The maximal exponent which satisfies the rate requirement of 4.1854 bit for an error-free "transmission" using the biological codon frequency is found to be ≈ 0.26. At the same exponent, the optimized "channel" capacity is 4.2586 bit. It can also be seen that the capacity curve is very close to the one found by using the biological codon distribution. This indicates that the biological probability distribution is almost optimally "chosen" to achieving the capacity of the "channel".

It is not surprising that the exponent is not one, since the matrix was obtained comparing five different vertebrate DNAs, the times corresponding to time spans between 40 M − 350 M years. However, the exponent is not extremely small, which indicates that the matrix is at least roughly in agreement with information-theoretic calculations. One may also see this as an argument to recompute the matrix using the obtained coefficient.

To see how well the biological and the optimized codon distributions agree, we applied the Kull-back–Leibler divergence (D_{KL}) and obtained a value of 0.0926 bit, which is not a very small difference (comparable with the D_{KL} of two Gaussians of equal mean and a variance differing by a factor of two) but still, similarities are obvious. Both of the probability distributions satisfy the rate requirement of 4.1875 bit. In addition, the distribution among synonymous codons is very similar. To mention one example, codons encoding Alanine (A) in decreasing order of abundance, is GCC, GCT, GCA, and GCG, for both the biological and the capacity-achieving distributions.
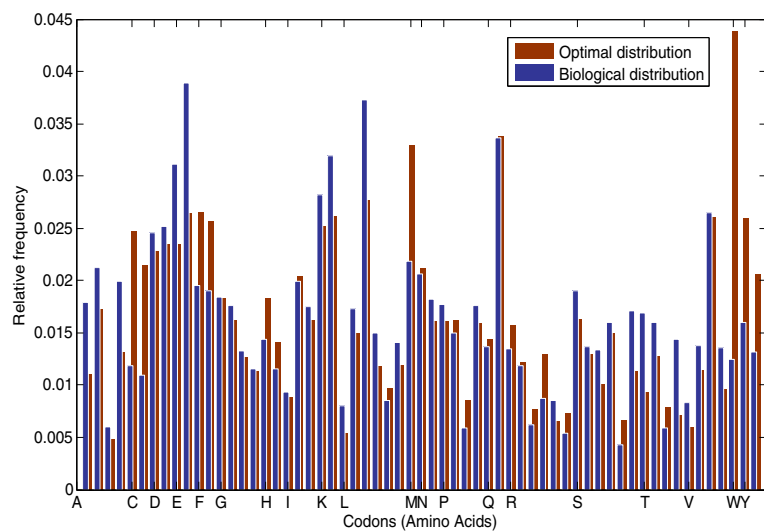
Figure 1 Probability distribution of codons (Biological and Optimal). The optimum codon distribution to maximize mutual information and the biological distribution of codons in the five vertebrates. Consecutive bins indicate that the codons belong to the same encoded amino acid (one letter symbol). The synonymous codons are arranged alphabetically.

Conclusion

From the so-called empirical codon substitution matrix (ECM), a mutation probability matrix, we derived the capacity when regarding the matrix as a communication channel. We found that an exponent of 0.26 would lead to a capacity of 4.1875 bit that is at least required to preserve the genetic information represented by the 20 amino acids encoded by 64 codons. Additionally, for the desired channel capacity, we have presented the optimal codon distribution found by searching the distribution that maximizes

Table 1 Biological codon relative frequency

	Codon	Frequency	Codon	Frequency	Codon	Frequency	Codon	Frequency	
T	TTT	0.0191	TCT	0.0171	TAT	0.0132	TGT	0.0110	T
	TTC	0.0196	TCC	0.0160	TAC	0.0160	TGC	0.0119	C
	TTA	0.0085	TCA	0.0133	TAA	0.0003	TGA	0.0003	A
	TTG	0.0141	TCG	0.0043	TAG	0.0001	TGG	0.0125	G
C	CTT	0.0150	CCT	0.0176	CAT	0.0116	CGT	0.0054	T
	CTC	0.0173	CCC	0.0150	CAC	0.0144	CGC	0.0087	C
	CTA	0.0080	CCA	0.0178	CAA	0.0137	CGA	0.0062	A
	CTG	0.0373	CCG	0.0059	CAG	0.0337	CGG	0.0085	G
A	ATT	0.0175	ACT	0.0144	AAT	0.0182	AGT	0.0136	T
	ATC	0.0200	ACC	0.0160	AAC	0.0206	AGC	0.0191	C
	ATA	0.0094	ACA	0.0169	AAA	0.0282	AGA	0.0135	A
	ATG	0.0219	ACG	0.0059	AAG	0.0319	AGG	0.0118	G
G	GTT	0.0136	GCT	0.0200	GAT	0.0252	GGT	0.0115	T
	GTC	0.0138	GCC	0.0213	GAC	0.0246	GGC	0.0176	C
	GTA	0.0084	GCA	0.0179	GAA	0.0311	GGA	0.0184	A
	GTG	0.0265	GCG	0.0060	GAG	0.0389	GGG	0.0133	G
	T		C		A		G		

The codon relative frequency of the five vertebrates genomes (human, mouse, chicken, frog, and zebrafish) from the data presented by Schneider A., Cannarozzi G., and Gonnet G. [4].

Table 2 Calculated codon relative frequency

	Codon	Frequency	Codon	Frequency	Codon	Frequency	Codon	Frequency	
T	TTT	0.0257	TCT	0.0113	TAT	0.0207	TGT	0.0215	T
	TTC	0.0264	TCC	0.0150	TAC	0.0260	TGC	0.0247	C
	TTA	0.0097	TCA	0.0100	TAA	*	TGA	*	A
	TTG	0.0119	TCG	0.0066	TAG	*	TGG	0.0439	G
C	CTT	0.0118	CCT	0.0159	CAT	0.0141	CGT	0.0073	T
	CTC	0.0150	CCC	0.0162	CAC	0.0183	CGC	0.0129	C
	CTA	0.0054	CCA	0.0161	CAA	0.0144	CGA	0.0077	A
	CTG	0.0277	CCG	0.0085	CAG	0.0337	CGG	0.0065	G
A	ATT	0.0162	ACT	0.0071	AAT	0.0160	AGT	0.0130	T
	ATC	0.0205	ACC	0.0128	AAC	0.0212	AGC	0.0163	C
	ATA	0.0088	ACA	0.0093	AAA	0.0251	AGA	0.0157	A
	ATG	0.0330	ACG	0.0079	AAG	0.0261	AGG	0.0122	G
G	GTT	0.0096	GCT	0.0132	GAT	0.0234	GGT	0.0114	T
	GTC	0.0114	GCC	0.0172	GAC	0.0228	GGC	0.0162	C
	GTA	0.0060	GCA	0.0110	GAA	0.0235	GGA	0.0183	A
	GTG	0.0260	GCG	0.0048	GAG	0.0263	GGG	0.0126	G
	T		C		A		G		

The codon relative frequency that maximizes the mutual information between input and output and yielding a capacity close to what is required for preserving the information content of amino acids. An exponential factor of 0.26 is applied to the ECM matrix.

the mutual information starting from a random initialization. A comparison of the biological distribution with the optimized codon distribution shows that the two distributions are not too similar. However, the biological distribution is not too far from the capacity-achieving distribution in terms of "channel" capacity, which indicates that the biological distribution is well "chosen". Additionally, the optimal codon distribution has preserved the relative abundance of synonymous codons. We concluded that the ECM as a channel is not too far from what would be expected following information theoretic arguments although it was derived from 5 different species.
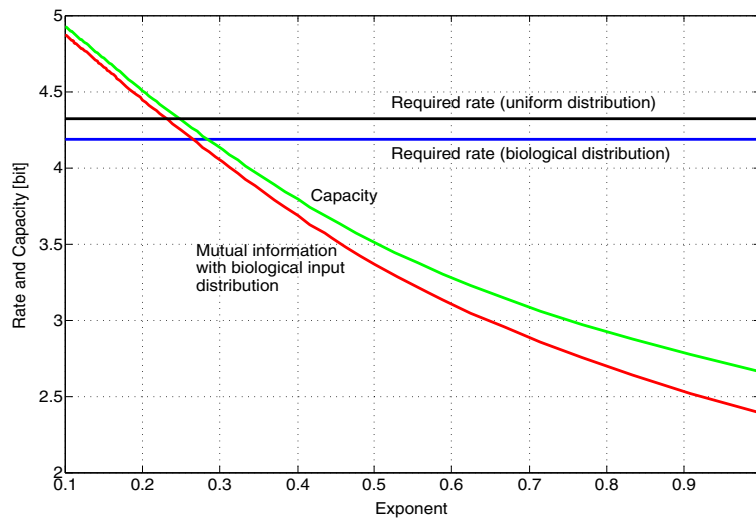

Figure 2 Capacity as a function of an exponential factor. The required rate for error-free transmission and the achievable capacity are plotted as a function of the exponent of the ECM matrix.

References

1. Dayhoff MO, Schwartz RM, Orcutt BC: **A model of evolutionary change in proteins.** *Natl Biomed Res Found, Wash, D.C.* 1978, **5**:345–352.
2. Henikoff S, Henikoff J: **Amino acid substitution matrices from protein blocks.** *Proc Natl Acad Sci USA* 1992, **89**:10915–10919.
3. Whelan S, Goldman N: **General empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach.** *Mol Biol Evol* 2001, **18**:691–699.
4. Schneider A, Cannarozzi G, Gonnet G: **Empirical codon substitution matrix.** *BMC Bioinformatics* 2005, **6**:134. [http://www.biomedcentral.com/1471-2105/6/134]
5. Yockey HP: **An application of information theory to the central dogma and the sequence hypothesis.** *J Theoretical Biol* 1974, **46**(2):369–406.
6. Yockey H: *Information Theory and Molecular Biology.* Cambridge, UK: Cambridge University Press; 1992.
7. Gong L, Bouaynaya N, Schonfeld D: **Information-theoretic model of evolution over protein communication channel.** *Comput Biol Bioinform, IEEE/ACM Trans* 2011, **8**:143–151.
8. Arimoto S: **An algorithm for computing the capacity of arbitrary discrete memoryless channels.** *Inform Theory, IEEE Trans* 1972, **18**:14–20.
9. Blahut R: **Computation of channel capacity and rate-distortion functions.** *Inform Theory, IEEE Trans* 1972, **18**(4):460–473.
10. Shannon CE: **A Mathematical theory of communication.** *Bell Syst Tech J* 1948, **27**:623–656. [http://cm.bell-labs.com/cm/ms/what/shannonday/shannon1948.pdf]
11. Cover TM, Thomas JA: *Elements of Information Theory.* New York, NY, USA: Wiley-Interscience; 1991.
12. King J, Jukes T: **Non-Darwinian evolution.** *Science (New York, NY)* 1969, **164**(3881):788.