Journal of Theoretical Biology 419 (2017) 227–237

Contents lists available at ScienceDirect

Journal of Theoretical Biology

journal homepage: www.elsevier.com/locate/yjtbi

The information capacity of the genetic code: Is the natural code optimal?

Ercan E. Kuruoglu[a,*], Peter F. Arndt[b]

[a] Institute of Information Science and Technologies, "A. Faedo", CNR, via G Moruzzi 1, 56124 Pisa, Italy
[b] Max Planck Institute for Molecular Genetics, Department of Computational Molecular Biology, Ihnestr. 63/73, 14195 Berlin, Germany

ARTICLE INFO

Keywords:
Genetic code
DNA
Information capacity
Shannon theory
Information theory

ABSTRACT

We envision the molecular evolution process as an information transfer process and provide a quantitative measure for information preservation in terms of the channel capacity according to the channel coding theorem of Shannon. We calculate Information capacities of DNA on the nucleotide (for non-coding DNA) and the amino acid (for coding DNA) level using various substitution models. We extend our results on coding DNA to a discussion about the optimality of the natural codon-amino acid code. We provide the results of an adaptive search algorithm in the code domain and demonstrate the existence of a large number of genetic codes with higher information capacity. Our results support the hypothesis of an ancient extension from a 2-nucleotide codon to the current 3-nucleotide codon code to encode the various amino acids.

1. Introduction

The fundamental biochemical processes in the cell such as replication, transcription, translation as well as cell signalling can be envisioned as information transfer processes. For some of these processes there is an original information carrying message stored in a biological entity (the DNA) that needs to be transferred to following generations through a noisy medium characterised by mutations. In the end the coding part of the DNA needs to be decoded to a protein, i.e the biological message which is originally stored in DNA needs to be transcribed into RNA and then translated into an amino acid sequence, two processes which might cause errors as well.

The paradigm of information transfer in biological systems brings into mind an analogy with communication systems (Fig. 1) where the message is coded into a waveform or a signal which carries the information coded in a way that it is compact, to save on material and energy, and robust to noise to prevent loss of information. The information carrying signal then is transferred over the noisy channel to be received at a receiver and decoded to recover the information.

This analogy was established by several researchers in the past in works as early as Jukes and Gatlin (1971), Yockey (1978), Román-Roldán et al. (1996), Battail (2004) and Konopka (2006). A key element of the analogy is the ability to quantify the information which is provided by the *entropy* as an information measure (Shannon, 1948). Numerous publications in the literature have studied the entropy of the DNA (Schneider and Spouge, 1997), across the species, at protein binding sites (Schneider, 2000, 2010), etc. The reader is referred to the paper by Fabris (2009) for a critical review and summary of earlier work and formulation of the information theory framework for various related problems. Some other works study the problem from purely coding theory point of view and try to discover hidden coding structures (May et al., 2004; Battail, 2004). Only a few works (Gong et al., 2011; Balado, 2013), however, attempted at a full analysis of the information transfer processes in the genome such as protein coding, to derive its fundamental limits.

Calculation of the fundamental limits of transfer of information is very important for the understanding of biological evolution over generations as well as the functioning of biological processes to decode the information stored in DNA. In particular, it can tell us the expected time or number of generations after which vital information about an organism would be lost during molecular evolution. It can also provide us insight into understanding the existing natural genetic (codon-amino acid) code and where it stands among all possible codes, in particular, whether nature tried to optimize the information capacity in choosing the natural code among a very large number of possible codes.

Although various previous publications build on the communications system analogy, most fail to address this problem, partly due to the over-idealisation of the analogy. In a typical communication system the messages are encoded and transmitted over noisy channels which are to be received, decoded and reconstructed as close as possible to the original message. It must be underlined that a full analogy with a communication system fails in the sense that the encoder is lacking in a biological system. In the case of protein coding, the decoded message is not a DNA but an amino acid sequence. In this case, one can at best talk of a hypothetical information source already coded in the form of a nucleotide sequence.

[*] Corresponding author.
E-mail address: ercan.kuruoglu@isti.cnr.it (E.E. Kuruoglu).

http://dx.doi.org/10.1016/j.jtbi.2017.01.046
Received 30 November 2016; Received in revised form 25 January 2017; Accepted 31 January 2017
Available online 03 February 2017

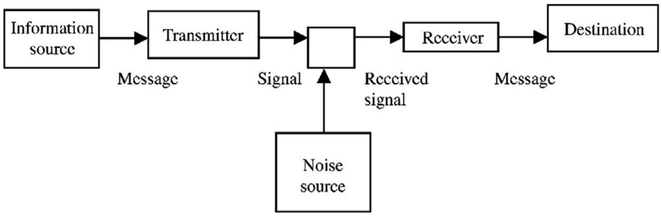

Fig. 1. A generic communications system.

In this article, utilizing the Coding Theory of Shannon, we develop theoretical limits of information preservation in non-coding and amino acid coding DNA in terms of the channel capacity. The channel noise is characterised by various mutation models widely accepted in the literature. The quantification of the information preservation capacity brings us to the discussion of the optimality of the natural genetic (codon-amino acid) code. This question was posed in the past by several researchers but the analyses were not done in terms of channel capacity. Furthermore, considering other possible codes only a very limited part of the entire space of codon-amino acid codes were explored. With this publication, we propose an “intelligent” search algorithm optimizing the channel capacity to find an optimal genetic code and to understand where the natural code stands with respect to an optimal code.

The rest of this article is organised as follows: the next section provides the fundamentals of entropy as a measure of information and of Shannon's coding theory and define channel capacity. We give channel capacity results on non-coding DNA and protein coding DNA in Section 2.2 and 2.3, respectively. The optimality of the natural codon-amino acid encoder is studied in Section 3. Conclusions and future research directions are provided in Section 4.

2. Methods

2.1. Information capacity

As in previous works on application of information theory in biology, we quantify (the lack of) information with entropy, following the definition of Shannon (1948):

$$H(p) = -\sum_i p_i \log_2 p_i, \tag{1}$$

where p_i is the probability of the i-th source symbol in the dictionary of possible symbols. As an example: for the observed human nucleotides distribution of $p_{[A,C,G,T]} = [0.29\ 0.21\ 0.21\ 0.29]$ (Yamagishi and Shimabukuro, 2008), the entropy is calculated to be $H(p_{[A,C,G,T]}) = 1.9815 < 2$. If the nucleotides were uniformly distributed, the entropy would have achieved the highest value of 2 for a dictionary of size 4. Similarly, the entropy of the codon distribution in humans is $H(p_{codons}) = 5.7936 < 3 \times H(p_{[A,C,G,T]}) = 5.9445$ using the frequencies reported in Nei and Kumar (2000). If all codons were equiprobably distributed it would have achieved the maximum value of 6. The fact that the entropy of codons is less than 3 times the entropy of nucleotides indicates a statistical dependency between the nucleotides in the codon.

Referring back to Fig. 1, the capacity of a channel is defined as the maximum of the mutual information between the input and the output of the channel.

$$C = \max_{p_X} I(X;\,Y) = \max_{p_X}(H(Y) - H(Y|X)) = \max_{p_X} \sum_{x,y} p(x,\,y)\log\frac{p(x,\,y)}{p(x)p(y)} \tag{2}$$

where $H(Y|X)$ is the conditional entropy of the output Y, given input X and the maximum is taken over all possible input distributions p_X. The Channel Capacity provides a measure of the maximum information one can transmit over a channel, the channel being characterised by $p(Y|X) = p(X,\,Y)p(X)$, the distribution of the noise in the channel.

The analytic calculation of the Channel Capacity is not easy other than for a limited number of special cases such as the Gaussian channel, binary symmetric channel and binary erasure channel (Cover and Thomas, 2005). However, a numerical algorithm exists for calculating the channel capacity in the other cases, which is called the Blahut-Arimoto algorithm (Blahut, 1972; Arimoto, 1972). The Blahut-Arimoto algorithm searches iteratively the optimal input distribution leading to the highest mutual information between the input and the output, which is a convex optimisation problem.

A communication channel is characterised by the noise in the channel. In the case of the DNA channel, the noise is generated by mutations. Mutations can be insertions, deletions or single nucleotide substitutions. In our analyses we consider only substitutions since they are the prevalent source of errors. We consider the non-coding DNA channel and coding DNA channel, which also includes the translation into amino acids, separately.

2.2. Non-coding DNA

We first calculate the information capacity for non-coding DNA. In this case, the nucleotides are considered as independent messages and the communication has a rate of 2 bits due to the four letter alphabet. For the nucleotide channel, various substitution models have been proposed in the literature. The simplest such model is the Jukes-Cantor model, which assumes the same probability of error or mutation rate for each nucleotide (Jukes, 1969). Hence, the substitution matrix is characterised with only one parameter, the nucleotide substitution rate q. The Jukes-Cantor rate matrix is given in

$$Q_{JC} = \begin{bmatrix} -3q & q & q & q \\ q & -3q & q & q \\ q & q & -3q & q \\ q & q & q & -3q \end{bmatrix} \tag{3}$$

where the row and column indices are $A,\,C,\,G,\,T$. Then, the transition probability matrix $P(Y|X)$ for a finite time interval t can be obtained as (Nei and Kumar, 2000)

$$P_{JC} = \exp(Q_{JC}t) = \begin{bmatrix} 1-3p & p & p & p \\ p & 1-3p & p & p \\ p & p & 1-3p & p \\ p & p & p & 1-3p \end{bmatrix} \tag{4}$$

where $p = (1 - \exp(-4qt))/4$. For m generations we have $P(Y(m)|X) = P(Y|X)^m$. From (2), the channel capacity after m generations or m cascaded channels in Fig. 1 is

$$C_m = \max_p I(X;\,Y(m)) = \max_p[H(Y(m)) - H(Y(m)|X)] \tag{5}$$

Since the channel is symmetric, a uniform input X leads to a uniform output $Y(m)$. The first term is maximized for the uniform case and is simply $\log|\mathcal{X}|$, where $|\mathcal{X}|$ is the cardinality of X. The second term is independent of the input and corresponds to the entropy of a row of the substitution probability matrix (the entropy of all the rows are the same). Using these simplifying arguments, the capacity for each generation is calculated without the need for the Blahut-Arimoto algorithm.

The results are given in Fig. 2 which show the exponential decline of information capacity of the non-coding DNA channel with increasing number of generations. The results show clearly that information (capacity) vanishes exponentially over generations and that the time scale is given by the mutation rate.

In the biological context, the substitution rates for the so called transversions(purine-pyrimidine substitutions) and transitions(purine-purine or pyrimidine-pyrimidine substitutions) are observed to be different due to the different chemical properties of purines (Adenine and Guanine) and pyrimidines (Cytosine and Thymine). A substitution