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

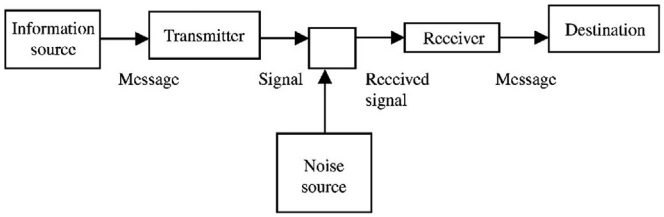


Fig. 1. A generic communications system.

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

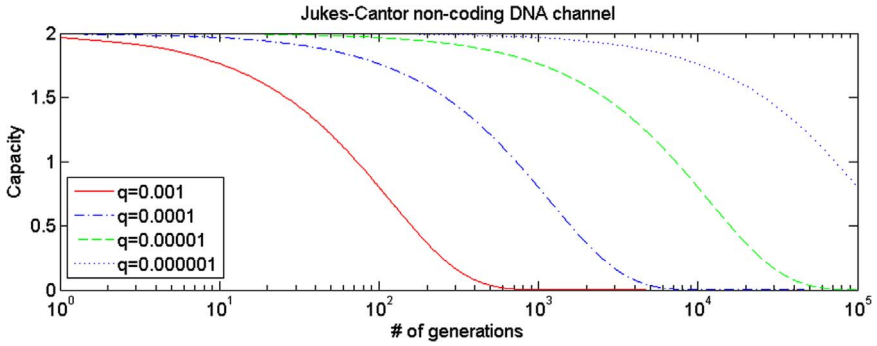


Fig. 2. Channel capacity for Jukes-Cantor non-coding DNA channel for various values of mutation rate in units of generation length.

model, which takes care of this effect, exists due to Kimura (1980). The Kimura rate matrix has two parameters and is given by

$$Q_{KM} = q\begin{bmatrix} -(2+K) & 1 & K & 1 \\ 1 & -(2+K) & 1 & K \\ K & 1 & -(2+K) & 1 \\ 1 & K & 1 & -(2+K) \end{bmatrix}. \quad (6)$$

Due to the symmetry of the matrix, we can invoke the same arguments as in the case of the Jukes-Cantor model and calculate the capacity from $C_m = \max_p I(X; Y(m)) = \max_p [H(Y(m)) - H(Y(m)|X)]$. The capacity curves are given in Fig. 3. The curve of the case K=1 corresponds to the Jukes-Cantor model and is included to provide a comparison. Increasing K indicates the dominance of transitions. In the limit of very large K, practically all substitutions are transitions and interchange between A and G or C and T, practically reducing the code to a 1-bit code rather than a 2-bit code.

These results show clearly the diversity in the capacity curves when one moves from equiprobable substitutions to unequal substitution rates for transitions and transversions.

The diversity in the capacity provided by Kimura model over Jukes-Cantor model might tempt one to look into more complex mutation models. We have therefore considered also the Felsenstein model (Felsenstein, 1981). The Felsenstein substitution rate matrix is given by:

$$Q_F = \begin{bmatrix} -(\pi_C+\pi_G+\pi_T) & \pi_C & \pi_G & \pi_T \\ \pi_A & -(\pi_A+\pi_G+\pi_T) & \pi_C & \pi_T \\ \pi_A & \pi_C & -(\pi_A+\pi_C+\pi_T) & \pi_T \\ \pi_A & \pi_C & \pi_G & -(\pi_A+\pi_C+\pi_G) \end{bmatrix} \quad (7)$$

where $\pi_A + \pi_C + \pi_G + \pi_T = 1$.

In this case, there is no symmetry anymore in the substitution matrix and there is no simplified way of calculating the capacity unlike in the Jukes-Cantor and Kimura cases. Therefore, the capacity is calculated using the Blahut-Arimoto algorithm. The obtained capacity curves for two different substitution vectors $[\pi_A\ \pi_C\ \pi_G\ \pi_T]$ are given in Fig. 4. As can be seen from the figure, although more diversity is obtained with the Felsenstein model, the difference in the capacity curves are limited.

Although for long, the non-coding part of DNA was seen as junk, now we have increasingly more knowledge about the function of parts of non-coding RNA as key regulators in translational and transcriptional control. In particular, studies have shown that long non-coding RNAs play a critical regulatory role in diverse cellular processes such as chromatin remodeling, transcription, post-transcriptional processing and intracellular trafficking (Ponting et al., 2009). The channel capacity of non-coding DNA can provide us an intuition on to what extend these functions can be preserved. It must be noted, however, that unlike the

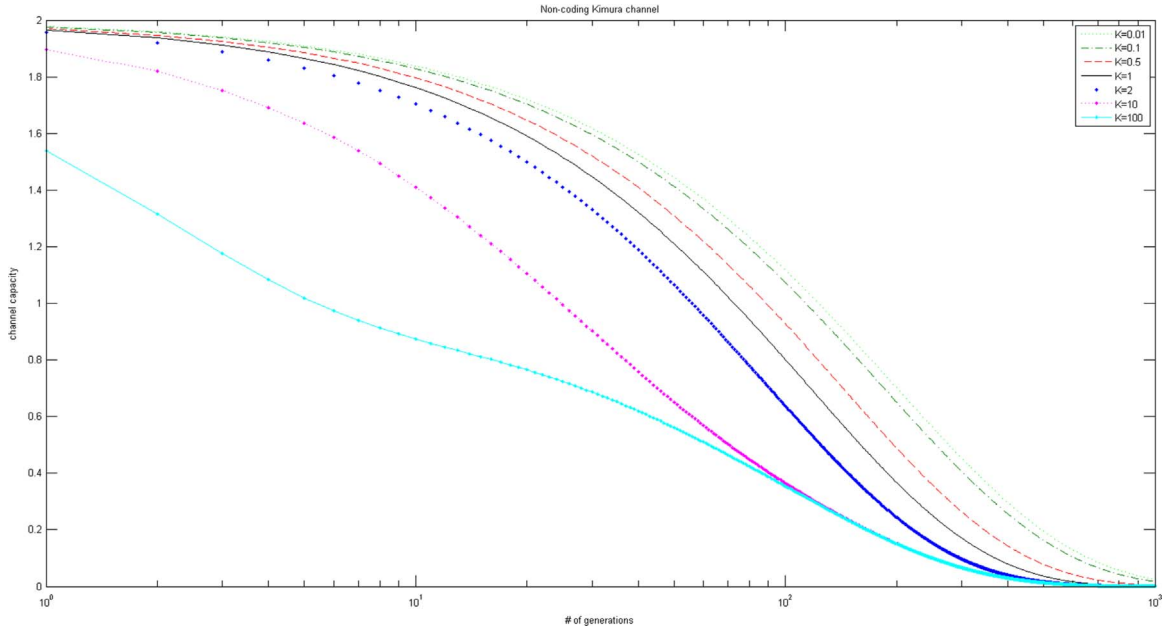


Fig. 3. Channel capacity for Kimura non-coding DNA channel for various values of transitions/transversions rate ratio K. q=0.001.

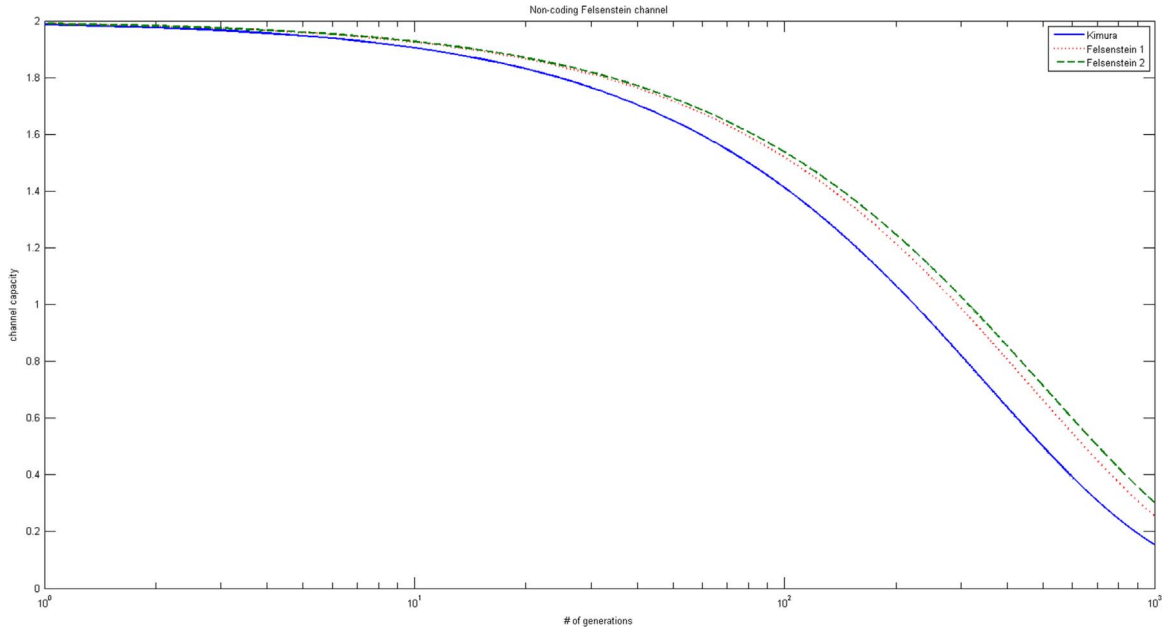
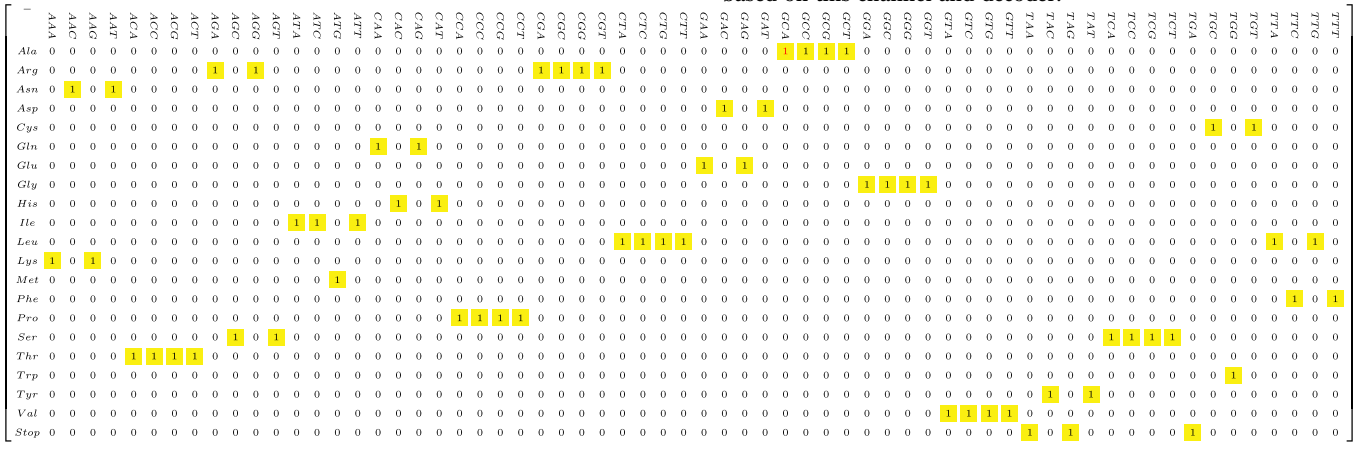


Fig. 4. Channel capacity for Felsenstein non-coding DNA channel for two values of mutation rate and comparison with Kimura channel. Kimura parameter: K=2, Felsenstein parameters: π_1 = [0.3 0.2 0.2 0.3], π_2 = [0.35 0.15 0.05 0.45].

coding DNA, these functions seem to be performed locally, that is the location information along the DNA sequence is important. The channel capacity calculations do not take such information into account and a second order/multivariate analysis is needed to calculate a location dependent capacity.

2.3. Coding DNA

In the case of non-coding DNA, the capacity analysis is straightforward since there is no obvious encoding structure. In the case of protein-coding DNA, considering the communication channel to have as input codons and as output amino acids, the presence of an encoder is clear. There are 64 codons (each codon being made of 3 nucleotides, $4^3 = 64$) which are mapped to 20 amino acids and some are used as stop markers. There is redundancy in the codon-amino acid mapping and this redundancy is used as an error correcting mechanism. The mapping between codons and amino acids is given in Fig. 5. This mapping can also be represented in matrix form as in Eq. (8).



(8)

One can define three different channels for this problem. The codon-codon channel, the codon-amino acid channel and the amino acid-amino acid channel. In Bouaynaya and Schonfeld (2007) and Gong et al. (2011), Bouyanaya et al. study the information transfer process between DNA and amino acids, underlining the breakdown of the communications system analogy and propose modelling the process with an amino acid-amino acid channel. That is, both the transmitted (X) and received (Y) signals are amino acids assuming a virtual protein source to DNA encoder. They characterised the communication channel using first the PAM250 matrix due to Dayhoff et al. (1978) and then by an amino acid transition matrix they constructed based on the assumption of Jukes-Cantor, equal-parameter nucleotide substitution matrix and they calculated the protein channel capacity.

Our approach differs from that of Bouyanaya et al. in that we underline that the mutations happen on the codons rather than on amino acids and therefore the codon substitution matrix needs to be propagated over generations, and not the amino acid substitution matrix. However, one should keep in mind that the ”meaning” of the message is in amino acids.

Using the Kimura nucleotide substitution model, we generate the corresponding codon (three-nucleotide) 64×64 substitution matrix. We propagate the message in the form of codons over generations and then decode the received codon to an amino acid and calculate the capacity based on this channel and decoder.

3. Results and discussion

It is curious that the natural genetic code (mapping) is not uniform. While most of the amino acids are coded by 2 different codons, some are coded by 6, 4, 3 or 1 codons. A natural question to ask is whether

1 / 2	T	C	A	G	3
T	14	11	10	20	T
	14	11	10	20	C
	11	11	13	20	A
	11	11	13	20	G
C	16	15	17	1	T
	16	15	17	1	C
	16	15	17	1	A
	16	15	17	1	G
A	19	9	3	4	T
	19	9	3	4	C
	21	6	12	7	A
	21	6	12	7	G
G	5	2	16	8	T
	5	2	16	8	C
	18	2	2	8	A
	18	2	2	8	G

Fig. 5. The natural genetic code (codon to amino acid map). 1: Alanine, 2: Arginine, 3: Asparagine, 4: Aspartate, 5: Cysteine, 6: Glutamate, 7: Glutamine, 8: Glycine, 9: Histidine, 10: Isoleucine, 11: Leucine, 12: Lysine, 13: Methionine, 14: Phenylalanine, 15: Proline, 16: Serine, 17: Threonine, 18: Tryptophan, 19: Tyrosine, 20: Valine, 21: STOP. We indicated the amino acids with numbers in the table to emphasize the fact that names are only labeling and should not affect our search for optimal codes in the sequel.

1 / 2	T	C	A	G	3
T	21	21	21	21	T
	21	21	21	21	C
	21	21	21	21	A
	21	21	21	21	G
C	21	21	21	21	T
	21	21	21	21	C
	21	21	21	21	A
	21	21	21	21	G
A	13	9	1	5	T
	14	10	2	6	C
	15	11	3	7	A
	16	12	4	8	G
G	21	21	17	21	T
	21	21	18	21	C
	21	21	19	21	A
	21	21	20	21	G

Fig. 6. The degenerate (extreme) genetic code (codon to amino acid map). 1: Alanine, 2: Arginine, 3: Asparagine, 4: Aspartate, 5: Cysteine, 6: Glutamate, 7: Glutamine, 8: Glycine, 9: Histidine, 10: Isoleucine, 11: Leucine, 12: Lysine, 13: Methionine, 14: Phenylalanine, 15: Proline, 16: Serine, 17: Threonine, 18: Tryptophan, 19: Tyrosine, 20: Valine, 21: STOP.

1 / 2	T	C	A	G	3
T	21	20	18	19	T
	21	20	17	19	C
	21	19	17	18	A
	21	20	17	18	G
C	16	15	12	14	T
	16	15	12	13	C
	15	14	11	13	A
	16	14	12	13	G
A	6	4	1	3	T
	5	4	1	2	C
	5	3	1	2	A
	5	4	2	3	G
G	11	10	7	8	T
	11	9	7	8	C
	10	9	6	7	A
	10	9	6	8	G

Fig. 7. The uniform-3 genetic code (codon to amino acid map). 1: Alanine, 2: Arginine, 3: Asparagine, 4: Aspartate, 5: Cysteine, 6: Glutamate, 7: Glutamine, 8: Glycine, 9: Histidine, 10: Isoleucine, 11: Leucine, 12: Lysine, 13: Methionine, 14: Phenylalanine, 15: Proline, 16: Serine, 17: Threonine, 18: Tryptophan, 19: Tyrosine, 20: Valine, 21: STOP.

1 / 2	T	C	A	G	3
T	21	19	15	17	T
	20	18	14	16	C
	20	18	14	16	A
	21	19	15	17	G
C	13	11	9	10	T
	12	11	9	10	C
	12	11	9	10	A
	13	11	9	10	G
A	4	3	1	2	T
	4	3	1	2	C
	4	3	1	2	A
	4	3	1	2	G
G	8	7	5	6	T
	8	7	5	6	C
	8	7	5	6	A
	8	7	5	6	G

Fig. 8. The uniform-42 genetic code (codon to amino acid map). 1: Alanine, 2: Arginine, 3: Asparagine, 4: Aspartate, 5: Cysteine, 6: Glutamate, 7: Glutamine, 8: Glycine, 9: Histidine, 10: Isoleucine, 11: Leucine, 12: Lysine, 13: Methionine, 14: Phenylalanine, 15: Proline, 16: Serine, 17: Threonine, 18: Tryptophan, 19: Tyrosine, 20: Valine, 21: STOP.

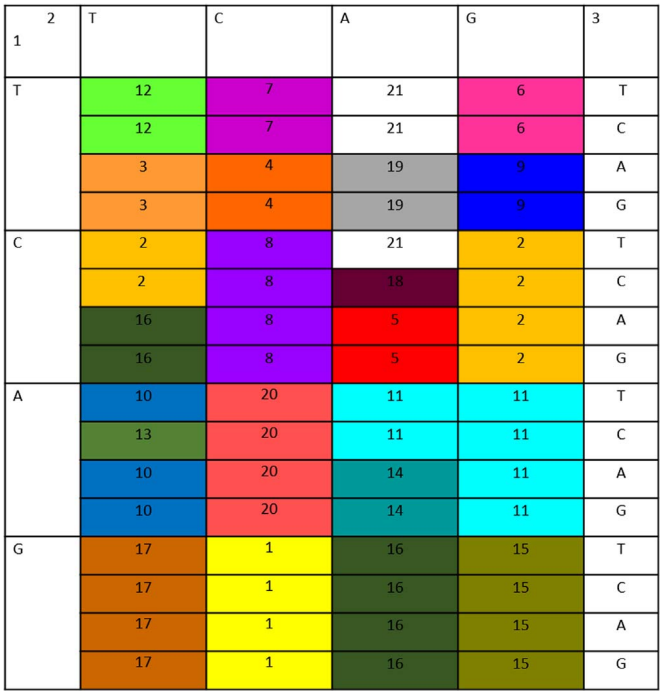


Fig. 9. The flipped natural genetic code (codon to amino acid map). 1: Alanine, 2: Arginine, 3: Asparagine, 4: Aspartate, 5: Cysteine, 6: Glutamate, 7: Glutamine, 8: Glycine, 9: Histidine, 10: Isoleucine, 11: Leucine, 12: Lysine, 13: Methionine, 14: Phenylalanine, 15: Proline, 16: Serine, 17: Threonine, 18: Tryptophan, 19: Tyrosine, 20: Valine, 21: STOP.

the natural genetic code is optimal in the information preservation, or channel capacity sense. To have an understanding of the space of possible codon-amino acid mappings, we have constructed a number of alternatives to the natural code:

1. an extreme-1 code where each amino acid is coded by only 1 codon and the remaining 44 codons are stop codons (Fig. 6).
2. a uniform code in which all amino acids are coded by 3 codons (and the stop codon by $64 - 20 \times 3 = 4$) which we will call the uniform 3 code (Fig. 7).
3. an almost uniform code in which the amino acids are coded by 4 or 2 codons, which we will call the uniform 4-2-code (Fig. 8).
4. a code obtained from the natural code by flipping C and G and A and T, for which transitions on the 3rd nucleotide would change the amino acid for 2-fold degenerate codons. We will call this the flipped natural code (Fig. 9).
5. similarly flipped version of the uniform 4-2 code (Fig. 10).

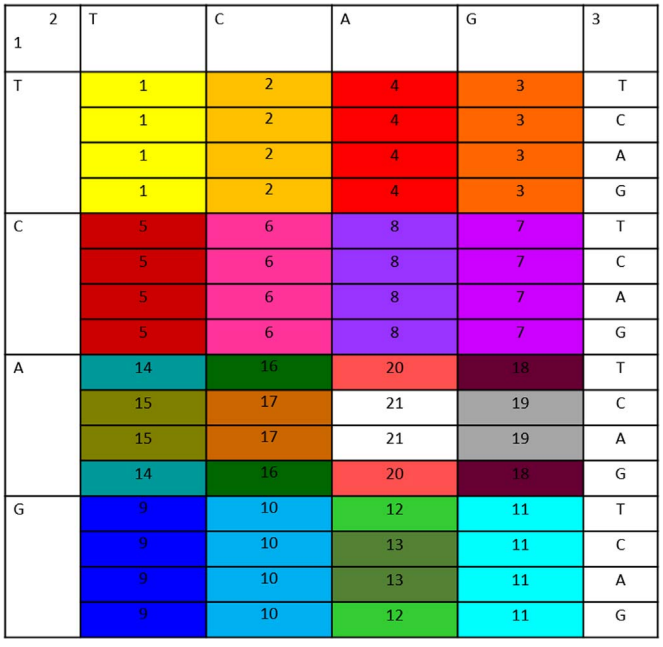


Fig. 10. The flipped uniform-42 genetic code (codon to amino acid map). 1: Alanine, 2: Arginine, 3: Asparagine, 4: Aspartate, 5: Cysteine, 6: Glutamate, 7: Glutamine, 8: Glycine, 9: Histidine, 10: Isoleucine, 11: Leucine, 12: Lysine, 13: Methionine, 14: Phenylalanine, 15: Proline, 16: Serine, 17: Threonine, 18: Tryptophan, 19: Tyrosine, 20: Valine, 21: STOP.

We have calculated the channel capacities for the natural amino acid code as well as the alternative codes using the Blahut-Arimoto algorithm, which are presented in Fig. 11. Several observations can be made on this figure: The channel capacity of the natural code is surpassed only by a uniform 4-2 code which has the same transitions-transversions structure as the natural code for $K > 1$. The extreme-1 code has the lowest channel capacity irrespective of the value of K. The flipped natural code has a higher channel capacity when $K < 1$, in which case transversions rather than transitions on the 3rd codon do not change the amino acid for 2 fold degenerate codons. The uniform-3 code has one of the lower channel capacity curves and surpass the natural code only for very small K. These observations tell us that the natural code favours a transitions dominant substitution model. It seems to be better than most alternative codes, however, falls slightly behind a uniform 4-2 code. This final observation emphasizes the fact that the natural code is not necessarily the optimal code at least in terms of channel capacity or information preservation or robustness to mutations.

These observations make us ask the question why the natural code was preferred to any other code. This question was asked before by several researchers including Crick who proposed the “frozen accident” model (Crick, 1968). The “frozen accident” model was questioned by various researchers in the literature who noted the “superiority” of the natural code to alternatives. For example, Freeland and Hurst (1998) generated randomly 1,000,000 different configurations and taking account of the mutation biases as in the Kimura model and using a mean square distance measure concluded that “the genetic code is one in a million”.

Other researchers use the *polar requirement*, a measure of hydrophobicity as the error measure and try to find/produce codes that minimize this cost function (e.g. Freeland et al., 2003). This group of work can be categorized as defending the “physicochemical theory”; the reader is referred to an interesting review by Tlusty (2010). Other work also attempt to define a “code fitness” measure (see Novozhilov et al., and references therein) based on distance functions weighted with substitution statistics and defined through the changes in synthesized proteins. The problem with such work is that substitution matrices such as PAM are employed which leads to a tautology (and the claim of the optimality of the natural code) since these matrices are the result of the natural genetic code (Di Giulio, 2001). For other references on the topic the reader is referred to Massey (2015).

Our approach is different from previous work in a number of aspects. Rather than using MSE (mean square error) on specific biochemical properties such as hydrophobicity, we use an information theory based measure which captures information on all statistics rather than only the second order statistics. The use of an MSE measure intrinsically makes a Gaussian distribution assumption which is not necessarily suggested by the nature of the data. The searches made in the literature seem to be random picks of codes from the space of possible codes such as in Freeland and Hurst (1998) which generated 1,000,000 different configurations but as noted in Santos and Monteagudo (2009), the explored code structures are rather rigid. Considering that there are $21^{64} \cong 4 \times 10^{84}$ configurations, this is a very limited sample to draw any conclusions from. In contrast, we propose

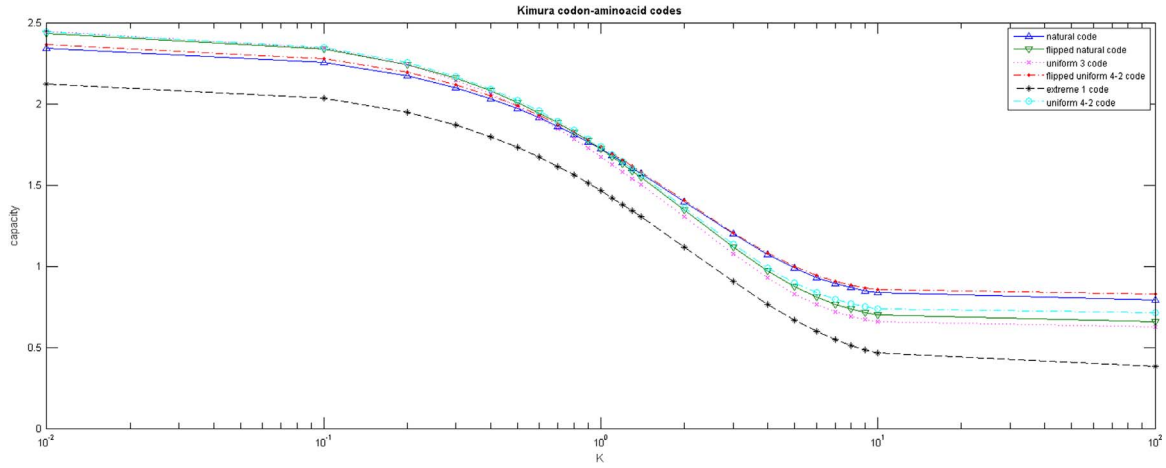


Fig. 11. Comparison of the Kimura channel capacities versus K (q=0.001) for various synthetic genetic codes and the natural genetic code. The values at 100th generation are plotted.

an intelligent search algorithm which learns through its search and searches at increasingly more promising parts of the space for solutions. The only other work which uses a learning intelligent algorithm is reported in Santos and Monteagudo (2009), however, they utilize a genetic algorithm rather than simulated annealing and the MSE measure on polar requirement as their cost function as opposed to the information theory based measure we use. The reader is referred to Sella and Ardell (2006) and the references therein for detailed accounts of past research on the "optimality" of the natural code.

Firstly, we start with a more realistic estimate of the available different configurations. We would like to partition m=64 labelled "items" (codons), to n=21 unlabelled non-empty "sets" (amino acids), unlabelled since we can rename the amino acids without loosing any biological meaning. This a classical problem in combinatorial mathematics and is called Stirling numbers of the 2nd kind. The number of configurations can be calculated using the formula:

$$S(m, n) = \frac{1}{n!}\sum_{i=0}^{n}(-1)^{i}C(n, i)(n-i)^{m} \tag{9}$$

where $C(n, i)$ is the combinatorial (n,i). We calculate $S(64, 21) = 2.9 \times 10^{64}$. We should also divide this by 4! since the order of A,C,G,T is arbitrary in constructing the matrix which gives 1.23×10^{63}. This number despite being much smaller than 21^{64}, is still too large a number to test all configurations.

We start by doing a limited search around the natural code searching all configurations of Hamming distance 2 to the natural code. We basically move a single 1 in the matrix in Eq. (8) to a new position in the same column (hence changing only two entries in the matrix), which amounts to remapping a codon to a new amino acid and calculate the channel capacity for all such generated new configurations. While doing this we ensure that all amino acids are encoded by at least one codon. Disregarding the case of rows with a single 1,

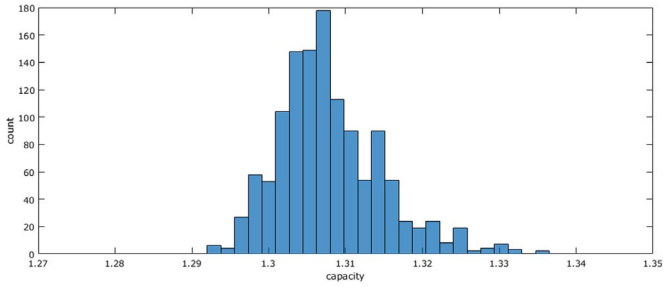


Fig. 12. Histogram of capacities (Kimura model) the genetic codes at Hamming distance 2 from the natural code. The natural code has capacity 1.3219. K = 2, q = 0.001.

1 / 2	T	C	A	G	3
T	14	11	10	20	T
	14	11	10	20	C
	11	11	13	20	A
	11	11	13	20	G
C	16	15	17	1	T
	16	15	17	1	C
	16	15	17	1	A
	16	15	17	1	G
A	19	9	3	4	T
	19	9	3	4	C
	21	6	12	7	A
	21	6	12	7	G
G	5	2	16	8	T
	5	2	16	8	C
	18	2	2	8	A
	18	2	2	8	G

Fig. 13. A genetic code 4-Hamming distance from the natural code (codon to amino acid map). 1: Alanine, 2: Arginine, 3: Asparagine, 4: Aspartate, 5: Cysteine, 6: Glutamate, 7: Glutamine, 8: Glycine, 9: Histidine, 10: Isoleucine, 11: Leucine, 12: Lysine, 13: Methionine, 14: Phenylalanine, 15: Proline, 16: Serine, 17: Threonine, 18: Tryptophan, 19: Tyrosine, 20: Valine, 21: STOP.

$62 \times 20 = 1240$ such configurations (Hamming distance 2 neighbours of the natural code). Below in Fig. 12, we provide the histogram of the capacities of all such configurations: The natural code is one of the best but not the best among its neighbours in terms of capacity. We can also construct a higher capacity code at Hamming distance 4 from the natural code with a simple observation. We have already shown the superiority of an 4-2 code above. When we look at the natural code, we see that the codons are mostly coded in groups of 4 or 2 to an amino acid with redundancies mostly at the third codon position and less at the first codon position, with the exceptions of Isoleucine (ATA, ATC, ATT), Methionine (ATG), Tryptophan (TGG) and the STOP codons (TAA, TAG, TGA). To keep the 4 and 2 redundancies, let's construct a neighbouring code to the natural code by moving TGA from STOP to Tryptophan and ATA from Isoleucine to Methionine as depicted in

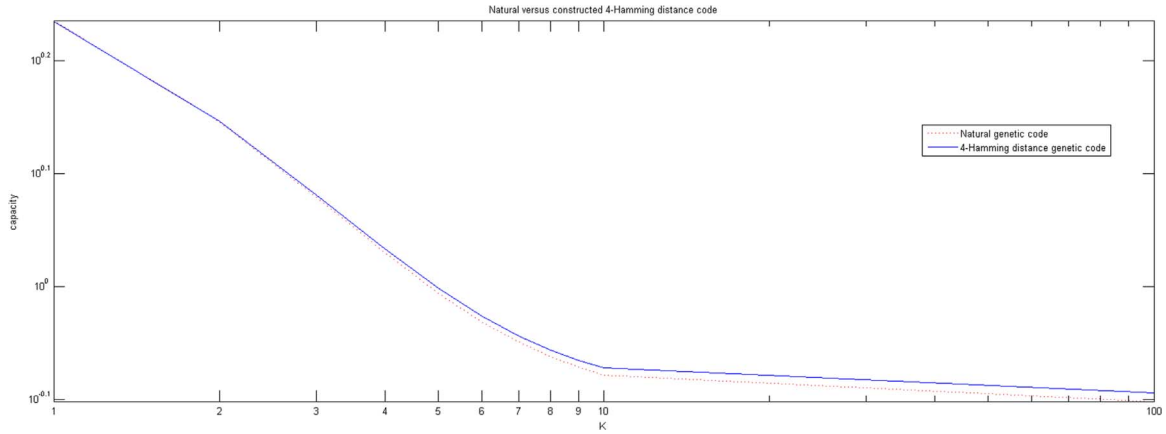


Fig. 14. Comparison of channel capacities for the natural genetic code and a constructed genetic code at Hamming distance 4 from the natural code. Kimura channel (K=2,q=0.001), 100 generations.

Fig. 13. The resulting code is at Hamming distance 4 from the natural code. As can be seen in Fig. 14, the channel capacity curve of this code is slightly above that of the natural code.

We can state that there are slightly more optimal codon-amino acid maps in the vicinity of the natural code. Either nature did not care to optimize the code even further or (more likely) there are hidden costs of some changes which we did not include in our considerations. For instance stop codons also play a vital role in the Nonsense Mediated Decay (NMD) pathway, having one less stop codon certainly affects the ability to detect nonsense errors during transcription. Further, it might be disadvantageous to have more than 1 codon coding for the start protein start (Met).

As mentioned above although several attempts exist to search for an optimal code, only random non-exhaustive searches have been made covering far less than a statistically meaningful space. The searches were not *intelligent* (that is not *learning* while progressing) leading to non-conclusive results. To search for a global optimum, we propose to use a non-convex optimisation algorithm, namely *Simulated Annealing* algorithm (Kirkpatrick et al., 1983), to do an intelligent search of the optimal code. The Simulated Annealing algorithm has had success in a wide variety application areas where the optimisation problem at hand is NP-hard, that is not solvable in polynomial time. These application areas include the traveling salesman problem, graph partitioning, scheduling in operations research, VLSI circuit design in electronics, optimal source coder design in telecommunications, etc (Kirkpatrick et al., 1983; Kuruoglu and Ayanoglu, 1993).

Simulated Annealing is motivated by experimental solid state physics where solids are first heated to a very high temperature and then cooled down slowly so that all electrons settle to their lowest energy states. The algorithm is motivated by the earlier ideas of Ulaby and Metropolis on chemical process modelling and is formulated by Kirkpatrick et al. in Kirkpatrick et al. (1983). Simulated Annealing proceeds with a series of random walks, namely Metropolis loops during which new configurations are proposed. If the new configuration leads to a better cost or energy (in our case the channel capacity), it is accepted. Unlike the steepest descent type of algorithms, simulated annealing occasionally accepts also worse configurations with certain probability given by Boltzmann statistics. This provides hill-climbing potential and the algorithm can avoid being stuck in local minima. The Boltzmann statistics provides the analogy with the modelling of the electron distribution in solid state physics. After each Metropolis loop, the temperature in the acceptance ratio is dropped, so less and less proposals with higher cost are accepted. It has been proved that if a logarithmic cooling schedule is applied the algorithm converges to the global optimum. However, a logarithmic cooling scheme can get infinitely slow and suboptimal schemes such as a geometric cooling scheme is applied. For detailed information on the simulated annealing algorithm, one is referred to van Laarhoven and Aarts (1987). A brief sketch of the algorithm is given below:

Simulated Annealing Algorithm

- Let $M = M_0$, where M_0 is the natural code matrix,
- While $T > T_{min}$,
  - $T \leftarrow T \times \alpha\alpha < 1$
  - Pick a random neighbour, $M_{new} \leftarrow N(M)$, where the neighbour set $N(.)$ includes all 2-Hamming distance codes from the code M
  - If $P(C(M), C(M_{new}), T) \geq$ random (0, 1), where $C(.)$ is the channel capacity and $P(.)$ is the Boltzmann function,
  - * then move to the new state $M \leftarrow M_{new}$
- Output: the final code *M and the channel capacity C.*

We have run the simulated annealing algorithm with geometric cooling scheme with a cooling coefficient of $\alpha = 0.99$. The starting configuration has been selected as the natural code. The new configurations are randomly selected by moving a 1 to a 0 in the amino acid-codon matrix. That is, changing the mapping of one codon from one amino acid to another amino acid making sure that there is at least one codon assigned to each amino acid. We have assumed uniform input distributions for the codons hence bypassing the Blahut-Arimoto algorithm. This choice was made since we do not have any prior information about the codon distribution and wanted to see the information preservation capability of the codes when no particular codon was emphasized by the nature.

Fig. 15 gives the evolution of capacity with progress of the simulated annealing algorithm to find the optimal code. It is interesting to note that the algorithm started with a strong drop in the capacity value (the algorithm accepted a worse code) and wild oscillations as expected in a simulated annealing run (the “temperature” is high in the beginning), then on the average improving the channel capacity by moving to “better” codes. Initially the changes are fast, reducing slowly and then saturating to significantly better codes or high capacity with small oscillations around the “near-optimal” codes. The initial drop of the capacity and the long time needed to recover the capacity in the run indicates that the natural code is already at a good point being better than most of its competitors although clearly being behind a large number of codes. The algorithm was rerun with different parameters such as lower initial temperature which lead to avoiding the initial drastic drop in the capacity and with smaller temperature coefficient leading to faster convergence. Various other starting points were chosen as well such as the “extreme” code or the “uniform 4-2″ code all leading to similar if not identical final result. The result of such a run starting with the extreme code is given in 16. It is interesting to note that in contrast to the case with the natural code as the starting point this Simulated Annealing run starts with a rapid increase in the

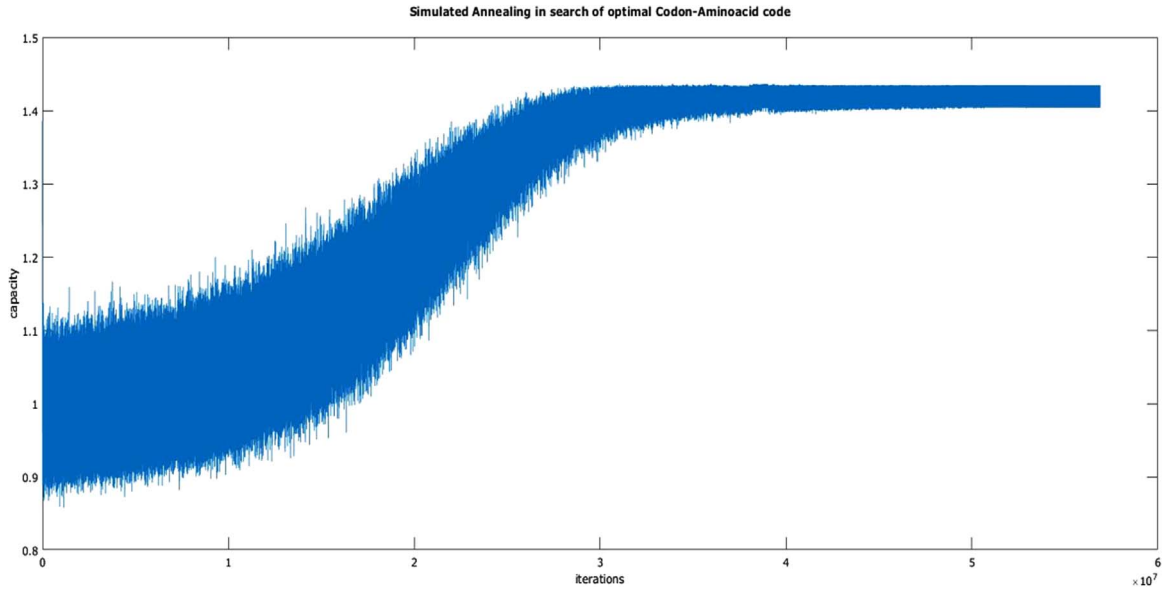


Fig. 15. The capacity of the code during the evolution of Simulated Annealing algorithm. The capacity value of the natural code is 1.38. Initial temperature is 0.1, temperature coefficient is 0.99.

capacity values as expected since the extreme code is a degenerate code with only one codon mapping to each amino acid.

The best configuration found in the simulations is given in Fig. 17 although some other codes exist with almost the same capacity value. It is very interesting to note that as in the case of the natural code, the codons producing the same amino acid are close in the table and have ambiguities in the nucleotides. The ambiguities in this optimal code are in the first (10 of them), second (8) and third (13) places. This is in contrast with the ambiguities seen in the natural code which are mostly at the third position (20) with some ambiguities also at the first position (2) but not at the second (0) position.

We provide a comparison of capacity profiles of this near optimal code with the natural code in Fig. 18. To give a scale of comparison, the capacity curves of the degenerate code (one codon synthesizing one amino acid) and a random 4-2 code are also plotted on the same figure. The figures show the channel capacity values at a certain number of generations for various values of the parameter K in the Kimura model corresponding the ratio of transitions/transversions. It can be seen that the near-optimal code obtained by the Simulated Annealing algorithm has significantly higher information capacity than the natural code. The difference is at the same scale as the difference between the natural code and the degenerate code and hence can be considered very significant. It is also worth noting that it is also significantly higher than the random 4-2 code discussed before constructed with ambiguities in the third place as in the case of the natural code.

These observations need a discussion on the biological significance. In particular, they underline clearly that the natural codon-amino acid code/map is far from being optimal although being better than most possible codes. The natural code can be “one in a million” (Freeland and Hurst, 1998); however, considering that there are more than 10^{63} possible configurations, being one in a million is not selective enough, it would mean still 10^{57} competitors. There are many other codes that have far better information preservation capabilities.

This observation may indirectly give support to three hypotheses.

1. that the genetic code co-evolved to a point that it would have been too disruptive to change anymore (Crick, 1968; Sella and Ardell, 2006), so its evolution was stopped prematurely.
2. that it is not completely an accidental code in that it is indeed an error-correcting code better than a large number of competitors (Ardell and Sella, 2002)
3. that at some point in the past the codons were composed of 2 nucleotides only and the third nucleotide was acquired afterwards. This may be the reason why the natural code does not seem to be optimized for 3-codons and that almost all redundancies are in the third position (Sella and Ardell, 1973; Santos and Monteagudo, 2009).

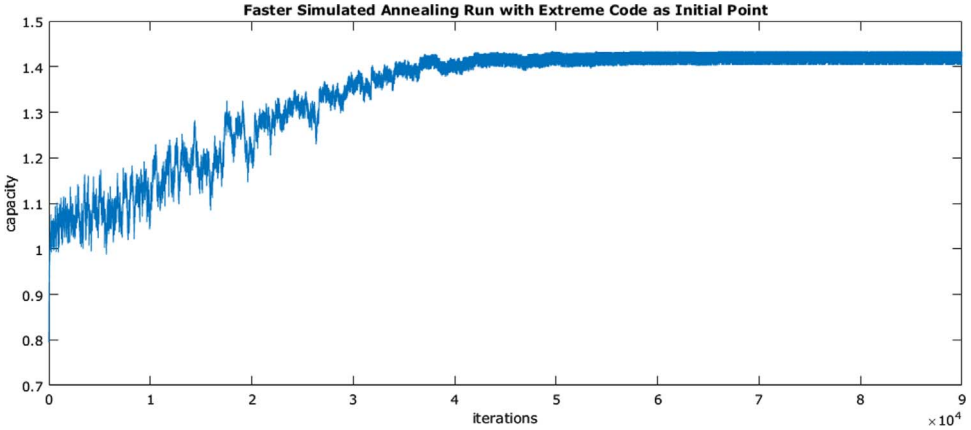


Fig. 16. The capacity of the code during the evolution of a fast run of Simulated Annealing algorithm with the extreme code as the initial code. The capacity value of the extreme code is 0.79. Initial temperature is 0.01, temperature coefficient is 0.95.

1 / 2	T	C	A	G	3
T	9	6	8	8	T
	9	6	15	15	C
	17	21	4	1	A
	17	21	4	1	G
C	9	6	8	8	T
	9	6	15	15	C
	17	21	4	1	A
	17	21	4	1	G
A	2	2	19	19	T
	14	14	13	13	C
	7	7	18	16	A
	7	7	18	16	G
G	10	10	5	5	T
	3	3	5	5	C
	11	20	12	12	A
	11	20	12	12	G

Fig. 17. The uniform-42 genetic code (codon to amino acid map). 1:Alanine, 2:Arginine, 3:Asparagine, 4:Aspartate, 5:Cysteine, 6:Glutamate, 7:Glutamine, 8:Glycine, 9:Histidine, 10:Isoleucine, 11:Leucine, 12:Lysine, 13:Methionine, 14:Phenylalanine, 15:Proline, 16:Serine, 17:Threonine, 18:Tryptophan, 19:Tyrosine, 20:Valine, 21:STOP.

Another biological problem to be discussed is whether the use of channel capacity as the optimality criterion of the protein code is justified. A higher capacity code definitely preserves the genetic information better over the generations; however, it also means less possibility for diversity. The error-correcting mechanism in the coding DNA is a sword with two edges. A completely preserved information would not allow diversity and selection.

The results we present are based on the Kimura substitution model. Other models such as the Felsenstein, HKY, TamuraNei model (Nei and Kumar, 2000) provide better approximations to the biological reality of nucleotide substitutions. They for instance can model stationary nucleotide distributions which are different from (1/4, 1/4, 1/4, 1/4) (Nei and Kumar, 2000). We have done preliminary simulation studies also with the Felsenstein model, which is one step higher in complexity compared to the Kimura model; however, we have observed that in terms of the relative channel capacities of alternative genetic codes, it did not lead to a significant changes.

In this paper, we have restricted our attention only to substitution type mutations. Although a full picture should include also insertion and deletion type mutations, this choice was motivated by two issues: the dominance of substitution type of mutations and the difficulty of calculating the channel capacity for deletion/insertion channels. This is indeed still an open problem in information/coding theory and is the subject of current research and only upper bounds for the channel capacity in these cases are known (Mitzenmacher, 2009; Fertonani et al., 2011).

Another point worth discussing is whether we could in principle also have used other cost functions besides the channel capacity. The simplest alternative is the mean squared error or other moments such as mean absolute deviations or nonlinear functions of the difference between the codons in subsequent generations. Although much simpler measures, these functions give only limited statistics. E.g. the mean squared distance provides us only the second order statistics of the errors. On the contrary, the channel capacity being based on mutual information carries information on all orders of statistics and hence is far more fundamental and informative.

4. Conclusions

In this paper, we have provided a complete modelling of the evolution process borrowing an analogy with communications, in terms of Shannon's coding theorems. Our model is different from previous work in that we consider a codon-amino acid channel rather than amino acid-amino acid or codon-codon channels as studied by researchers in the literature. We use the channel capacity as a measure of information preserving capability of the code and use it as a cost function to test the optimality of the natural protein (codon to amino acid) code. Given this cost function, we demonstrate the suboptimality of the natural code without any space for doubt. Its channel capacity is significantly below that of various other codes. Unlike previous work, we have extended our search space (close to 60 million tested configurations, that is almost 2 orders of magnitude higher than those reported in the literature) but more importantly we have done our search not “blindly” but “intelligently” using a non-convex learning/optimisation algorithm, namely Simulated Annealing. The method has indicated a large number of mappings different from the natural code and with redundancies in all three nucleotide positions while the natural code has redundancies mostly in the third place and never in the second place. This observation may be interpreted as a support for the hypothesis that once the codons were formed of 2-nucleotides only and that the third nucleotide was acquired later. The presented formulation, which places the information capacity as a measure of

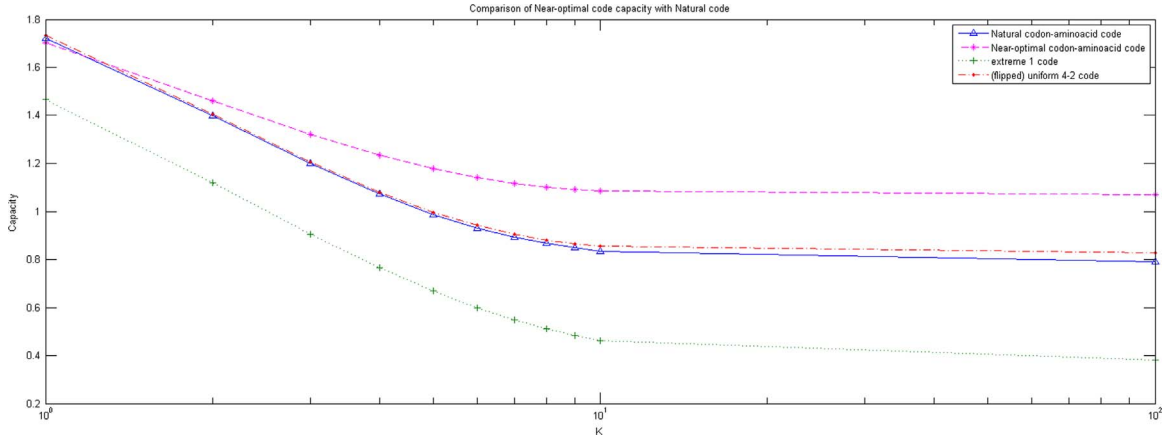


Fig. 18. Comparison of Channel capacity of Natural, Near-Optimal, Degenerate, Random 4-2 codes on Kimura codon-amino acid channel for various values of mutation rate at N generations.

the robustness of the genetic code, provides a mathematical framework for studying further biological questions.

Acknowledgements

This project was principally funded by the Alexander von Humboldt Foundation in the form of an Experienced Research Fellowship awarded to EE Kuruoglu. EE Kuruoglu also acknowledges partial support from CNR Short Term Mobility Program. The authors would like to thank M Vingron and A Bolshoy whose comments helped improve this work.

References

Ardell, D., Sella, G., 2002. No accident: genetic codes freeze in error-correcting patterns of the standard genetic code. Philos. Trans. R. Soc. B: Biol. Sci. 357 (1427), 1625–1642.
Arimoto, S., 1972. An algorithm for computing the capacity of arbitrary discrete memoryless channels. IEEE Trans. Inf. Theory 18 (1), 14–20.
Balado, F., 2013. Capacity of dna data embedding under substitution mutations. IEEE Trans. Inf. Theory 59 (2), 928–941.
Battail, G., 2004. An engineer's view on genetic information and biological evolution. Biosystems 76 (13), 279–290.
Battail, G., 2004. Can we explain the faithful communication of genetic information?. In: Siegel, P., Soljanin, E., Van Wijngaarden, A.J., Vasic, B. (Eds.), Advances in Information Recording, Vol. 8, American Mathematical Society: DIMACS-Series in Discrete Mathematics and Theoretical Computer Science, Ch. 10, pp. 79–103.
Blahut, R., 1972. Computation of channel capacity and rate-distortion functions. IEEE Trans. Inf. Theory 18 (4), 460–473.
Bouaynaya, N., Schonfeld, D., 2007. Protein communication system: evolution and genomic structure. Algorithmica NY. 48 (4), 375–397.
Cover, T., Thomas, J., 2005. Elements of Information Theory. John Wiley & Sons, Hoboken, New Jersey.
Crick, F., 1968. The origin of the genetic code. J. Mol. Biol. 38 (3), 367–379.
Dayhoff, B.C., Schwartz, M.O., Orcutt, R.M., 1978. A model of evolutionary change in proteins. Atlas Protein Seq. Struct. 5 (3), 345–352.
Di Giulio, M., 2001. The origin of the genetic code cannot be studied using measurements based on the pam matrix because this matrix reflects the code itself, making any such analyses tautologous. J. Theor. Biol. 208 (2), 141–144.
Fabris, F., 2009. Shannon information theory and molecular biology. J. Interdiscip. Math. 12 (1), 41–87.
Felsenstein, J., 1981. Evolutionary trees from dna sequences: a maximum likelihood approach. J. Mol. Evol. 17 (6), 368–376.
Fertonani, D., Duman, T., Erden, M., 2011. Bounds on the capacity of channels with insertions, deletions and substitutions. IEEE Trans. Commun. 59 (1), 2–6.
Freeland, S., Hurst, L., 1998. The genetic code is one in a million. J. Mol. Evol. 47 (3), 238–248.
Freeland, S., Wu, T., Keulmann, N., 2003. The case for an error minimizing standard genetic code. Orig. Life Evol. Biosph. 33 (4–5), 457–477.
Gong, L., Bouaynaya, N., Schonfeld, D., 2011. Information-theoretic model of evolution over protein communication channel. IEEE/ACM Trans. Comput. Biol. Bioinforma. 8 (1), 143–151.
Jukes, T., Gatlin, L., 1971. Recent studies concerning the coding mechanism. Prog. Nucleic Acid Res. Mol. Biol. 11, 303–350.
Jukes, T., 1969. Recent problems in the genetic code. Curr. Top. Microbiol. Immunol. 49, 178–219.
Kimura, M., 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J. Mol. Evol. 16 (2), 111–120.
Kirkpatrick, S., Gelatt, C.D., Vecchi, M.P., 1983. Optimization by simulated annealing. Science 220 (4598), 671–680.
Konopka, A., 2006. Information theories in molecular biology and genomics, eLS. Wiley http://dx.doi.org/10.1038/npg.els.0005927.
Kuruoglu, E., Ayanoglu, E., 1993. The design of finite-state machines for quantization using simulated annealing. In: Calderbank, R., Forney, G., Moayeri, N. (Eds.), Coding and Quantization: DIMACS/IEEE Workshop, October 19–21, 1992, American Mathematical Society: DIMACS-Series in Discrete Mathematics and Theoretical Computer Science, pp. 175–184.
Massey, S., 2015. Genetic code evolution reveals the neutral emergence of mutational robustness, and information as an evolutionary constraint. Life 5 (2), 1301–1332.
May, E., Vouk, M., Bitzer, D., Rosnick, D., 2004. An error-correcting code framework for genetic sequence analysis. J. Frankl. Inst. 341 (1–2), 89–109.
Mitzenmacher, M., 2009. A survey of results for deletion channels and related synchronization channels. Probab. Surv. 6 (1), 1–33.
Nei, M., Kumar, S., 2000. Molecular Evolution and Phylogenetics. Oxford University Press, Oxford, UK.
Novozhilov, A., Wolf, Y., Koonin, E. Evolution of the genetic code: Partial optimization of a random code for robustness to translation error in a rugged fitness landscape. Biology Direct 2.
Ponting, C., Oliver, P., Reik, W., 2009. Evolution and functions of long noncoding {RNAs}. Cell 136 (4), 629–641.
Román-Roldán, R., Bernaola-Galván, P., Oliver, J., 1996. Application of information theory to dna sequence analysis: a review. Pattern Recognit. 29 (7), 1187–1194.
Santos, J., Monteagudo, A., 2009. Genetic code optimality studied by means of simulated evolution and within the coevolution theory of the canonical code organization. Nat. Comput. 8 (4), 719–738.
Schneider, T., Spouge, J., 1997. Information content of individual genetic sequences. J. Theor. Biol. 189 (4), 427–441.
Schneider, T., 2000. Evolution of biological information. Nucleic Acids Res. 28 (14), 2794–2799.
Schneider, T., 2010. A brief review of molecular information theory. Nano Commun. Netw. 1 (3), 173–180.
Sella, G., Ardell, D., 1973. Possibilities for the evolution of the genetic code from a preceding form. Nature 246, 22–26.
Sella, G., Ardell, D., 2006. The coevolution of genes and genetic codes: crick's frozen accident revisited. J. Mol. Evol. 63 (3), 297–313.
Shannon, C., 1948. A mathematical theory of communication. Bell Syst. Tech. J. 27 (3), 379–423.
Tlusty, T., 2010. A colorful origin for the genetic code: information theory, statistical mechanics and the emergence of molecular codes. Phys. Life Rev. 7 (3), 362376.
van Laarhoven, P., Aarts, E., 1987. Simulated Annealing: Theory and Applications. Springer, Dordrecht, Holland.
Yamagishi, M.E.B., Shimabukuro, A.I., 2008. Nucleotide frequencies in human genome and fibonacci numbers. Bull. Math. Biol. 70 (3), 643–653.
Yockey, H., 1978. Can the central dogma be derived from information theory? J. Theor. Biol. 74 (1), 149–152.