The Information Capacity of the Genetic Code: is the Natural Code Optimal?

Ercan Engin Kuruoglu

Institute of Information Science and Technologies, "A. Faedo"-CNR, via G Moruzzi 1, 56124, Pisa, Italy. ercan.kuruoglu@isti.cnr.it

Peter F. Arndt

Max Planck Institute for Molecular Genetics, Department of Computational Molecular Biology, Ihnestr. 63/73, 14195, Berlin, Germany

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. Information capacities of both non-coding DNA and coding DNA are calculated using various mutation substitution models. We extend our results on coding DNA to a discussion about the optimality of the natural codon-aminoacid code. We provide the results of an intelligent search in the code domain and demonstrate the existence of a large number of genetic codes with higher information capacity. The results support the thesis of move from original 2-nucleotide codons to the current 3-nucleotide codons.

Keywords: genetic code, DNA, information capacity, Shannon theory, information theory.

[1]Corresponding author

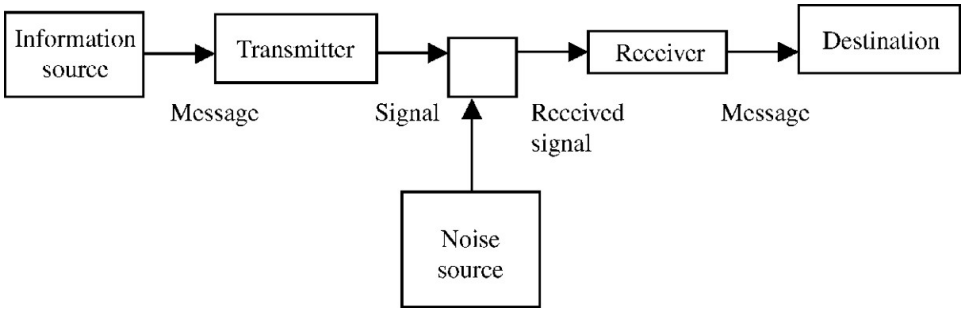


Figure 1: A generic communications system.

1. Introduction

The fundamental biochemical processes in the cell such as replication and translation as well as cell signalling can be envisioned as information transfer processes. In all such processes there is an original message stored in a biological apparatus (the DNA) that needs to be transferred through a noisy medium to another biological apparatus (the RNA polymerase). The DNA is stored in the nucleus of a cell and spontaneous mutations can change its sequence. In the example of translation, the biological message which is originally stored in DNA needs to be transcribed into RNA and then translated into proteins, two processes which might cause errors as well.

The paradigm of information transfer in biological systems brings into mind an analogy with communication systems (Figure 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 physical material, 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 obtain the information.

This analogy was established by several researchers in the past in works as early as [1, 2, 3]. A key element of the analogy is the ability to quantify the information which is provided by the entropy as an information measure [4]. Numerous publications in the literature have studied the entropy of the DNA [5], across the species, at protein binding sites [6, 7], etc. Very few works, however, did 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 in understanding the biological evolution over generations as well as the functioning of genomic processes. 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-aminoacid code) and where it stands among all possible codes. In particular, whether the nature tried to optimize the information capacity in choosing the natural code among a very large number of options.

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

In this article, utilizing the Coding Theory of Shannon, we develop theoretical limits of information preservation in non-coding DNA and aminoacid

coding 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-aminoacid) code. This question was posed several times but the analyses were not done in terms of channel capacity. With this publication, we propose an intelligent search algorithm optimising the channel capacity to find an optimal genetic code and to understand where the natural code stands with respect to the optimal code.

The rest of this article is organised as follows: the next section provides the fundamentals on entropy as a measure of information as the building stone of the model, Shannon's information and coding theory principles leading to channel capacity. We give channel capacity results on non-coding DNA and protein coding DNA in Section 3 and Section 4, respectively. The optimality of the natural codon-aminoacid encoder is studied in Section 5. Conclusions and future research directions are provided in Section 6.

2. Information Capacity

As in previous works on application of information theory in biology, we quantify (lack of) information with entropy, following the definition of Shannon [4]:

$$H(p) = -\sum_i p_i \log p_i. \tag{1}$$

As an example: for given human nucleotides distribution of $p_{[A,C,G,T]} = [0.29\ 0.21\ 0.21\ 0.29]$, the entropy is calculated to be $H(P) = 1.9815 < 2$. If the nucleotides were uniformly distributed, it would have achieved the highest value of 2. Similarly, the entropy of codon distribution in human is

$5.7936 < 3 \times H(p_{[A,C,G,T]}) = 5.9445$. If all the codons were equiprobably it would have achieved the highest value 6. The fact that the entropy of codons is less than 3 times the entropy of nucleotides indicates to a dependency structure between the nucleotides in the codon.

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

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

where $H(Y|X)$ is the conditional entropy of the output Y, given input X. The Channel Capacity provides a measure of the maximum information one can transmit over a channel, the channel being characterised by $p(Y|X)$, the distribution of the noise on the channel.

The calculation of Channel Capacity analytically is not easy other than for a limited number of special cases such as the Gaussian channel, binary symmetric channel and binary erasure channel [8]. However, a numerical algorithm exists for calculating the channel capacity in the other cases, which is called the Blahut-Arimoto Algorithm [9, 10]. The Blahut-Arimoto algorithm iteratively searches 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 noise in the channel. In the case of the DNA channel, the noise are the mutations. Mutations can be of type insertions, deletions or substitutions. In our analyses we consider mainly substitutions due to their dominance and the ease of work. We consider the non-coding DNA channel and coding DNA channel separately.

3. Non-Coding DNA

We first calculate the information capacity of the non-coding DNA. Only the nucleotides are considered as independent messages and the communication has 2-bit rate due to the existence of a four letter alphabet. For the nucleotide channels, various mutation models have been proposed before. The simplest such model is the Jukes-Cantor model which assumes the same probability of error or mutation rate for each nucleotide [11]. Hence, the substitution matrix is characterized 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 or substitution matrix for a finite time interval t can be obtained as ([12])

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

where $p = (1 - \exp(-4qt))/4$. From 2, the channel capacity after m generations or m cascaded channels in Figure 1 is

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

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

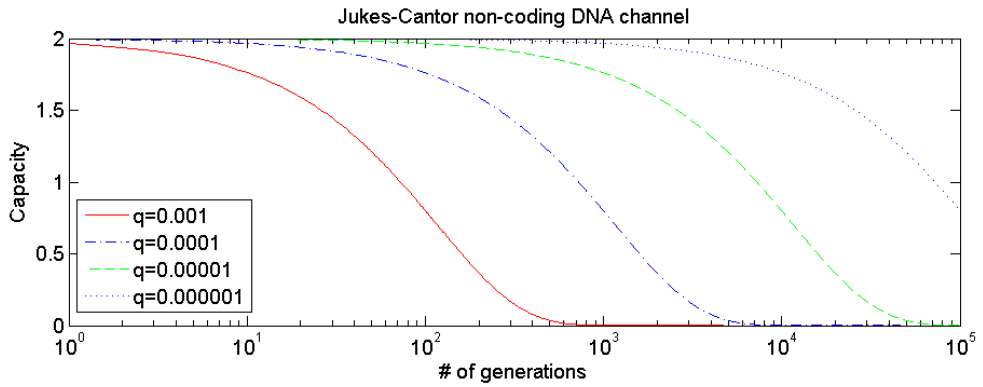


Figure 2: Channel capacity for Jukes-Cantor non-coding DNA channel for various values of mutation rate.

probability matrix (entropy of all the rows are the same). Using these simplifying arguments, the capacity for each generation is calculated without the need of using the Blahut-Arimoto algorithm.

The results are given in Figure 2 which show the exponential decline of information capacity of the non-coding DNA code with increasing number of generations. The curves potentially give us the information preservation limits of the DNA code over generations.

The results show clearly that information (capacity) vanishes exponentially over generations and that the time scale is given by the mutation rate. 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 DNA as key regulators in translational and transcriptional output. In particular, studies have shown that long non-coding dNAs play critical regulatory roles in diverse cellular processes such as chromatin remodeling, transcription, post-transcriptional processing and intracellular trafficking [13]. The channel capacity of non-coding DNA can provide us an intuition about how

far these functions can be preserved.

The channel capacity of non-coding DNA can be especially informative for viruses. It can help one predict in how many generations a virus would be dysfunctional due to the lack of an error correcting mechanism unlike coding DNA, and in the contrary sense it can also help understand the quick evolution into other viruses.

The Jukes-Cantor mutation model provides only an approximation to the actual mutation statistics since in the nature not all substitutions are equiprobable. The rates of substitutions of type transitions (purine-pyrimidine substitutions) and transversions (purine-purine or pyrimidine-pyrimidine substituons) are different due to the different chemical properties of purines (Adenine and Guanine) and pyrimidines (Cytosine and Thymine). A mutation substitution model with two parameters exists due to Kimura [14]. The Kimura substitution rate matrix is given by

$$Q_{km} = q \begin{bmatrix} -(2+K) & 1 & 1K & 1 \\ 1 & -(2+K) & 1 & K \\ K & 1 & -(2+K) & 1 \\ 1 & K & 1 & -(2+K) \end{bmatrix} \tag{6}$$

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 $C_m = \max_p I(X;Y(m)) = H(Y(m)) - H(Y(m)|X)$. The capacity curves are given in Figure 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 of type transitions and between A and G or C and T, practically reducing the code to a 1-bit code rather than a 2-bit code.

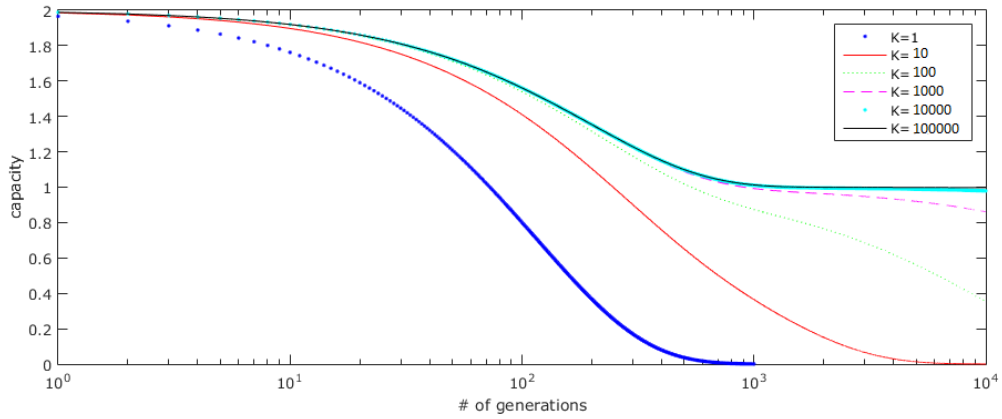


Figure 3: Channel capacity for Kimura non-coding DNA channel for various values of transitions/transversions rate ratio K.

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

The capacity gain with the diversity 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 [15]. 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 more symmetry in the substitution matrix and there is no simplified way of calculating the capacity unlike the Jukes-Cantor and Kimura cases. Therefore, the capacity is calculated using the Blahut-Arimoto algorithm. The obtained capacity curves for two different substi-

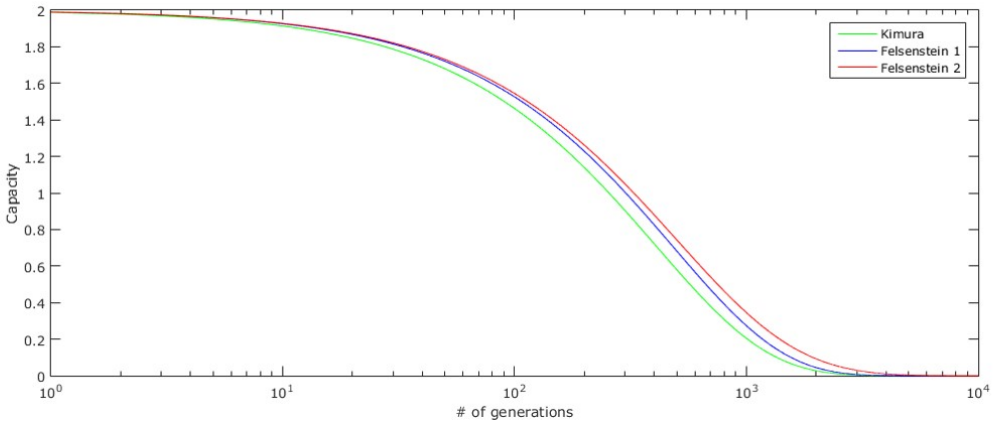


Figure 4: Channel capacity for Felsenstein non-coding DNA channel for two values of mutation rate and comparison with Kimura channel.

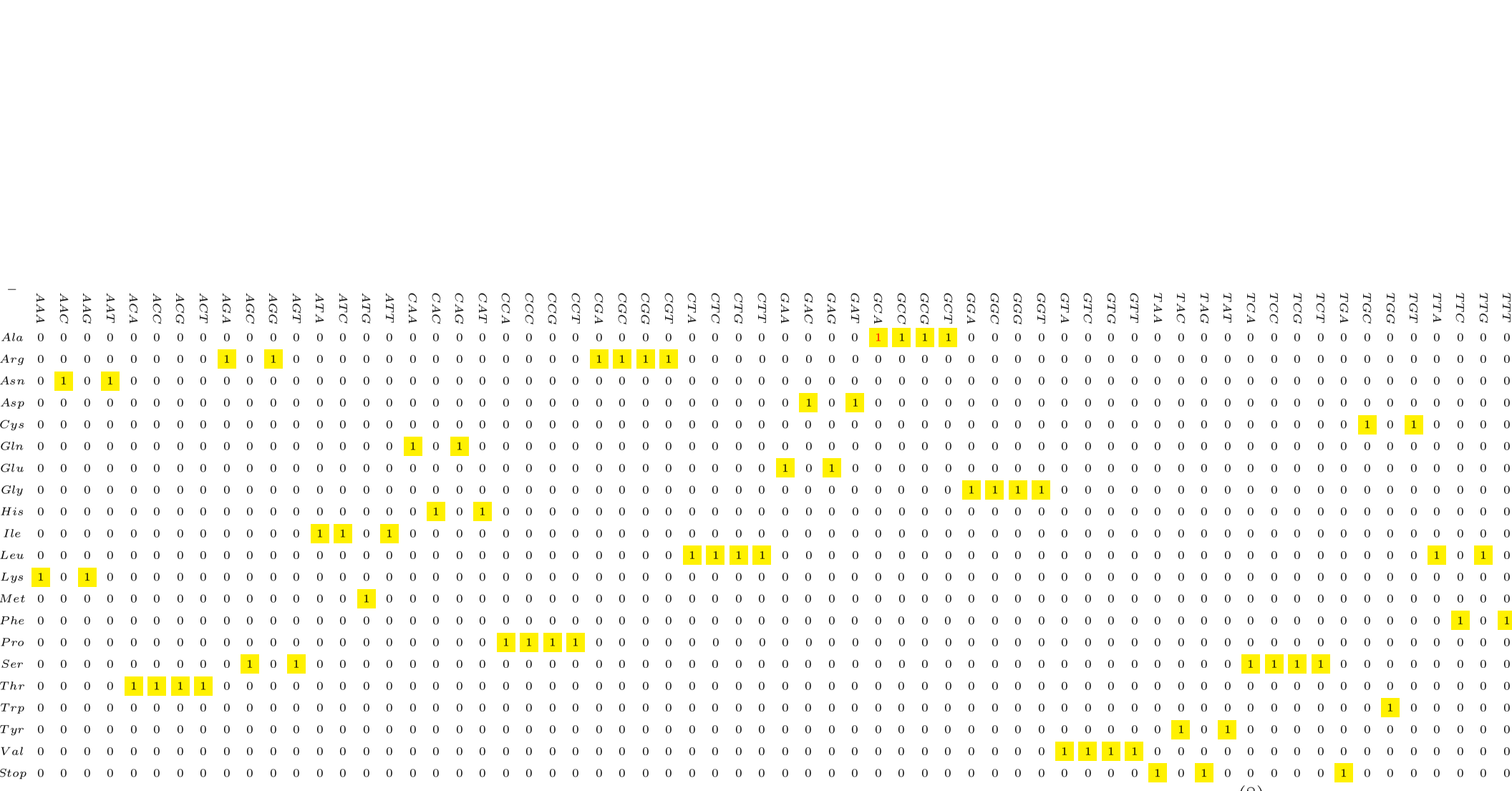
tution vectors $[\pi_A\ \pi_C\ \pi_G\ \pi_T]$ are given in Figure 4. As can be seen from the figure, although more diversity is obtained with the Felsenstein model, the difference in the capacity curves are limited.

4. 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 aminoacids, the presence of an encoder is clear. There are 64 codons (each codon being made of 3 nucleotides) which are mapped to 20 aminoacids and some are used as stop markers. There is redundancy in the codon-to-aminoacid map and this redundancy is used as an error correcting mechanism. The map between codons to aminoacids is given in Figure 5. This mapping can also be represented in matrix form as in Eq. 8.

1 2	T	C	A	G	3
T	14	11	10	20	T
	14	11	10	20	C
	11	11	10	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
	21	2	2	8	A
	18	2	2	8	G

Figure 5: The natural genetic code (codon to aminoacid 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 various aminoacids 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.


AAA AAC AAG AAT ACA ACC ACG ACT AGA AGC AGG AGT ATA ATC ATG ATT CAA CAC CAG CAT CCA CCC CCG CCT CGA CGC CGG CGT CTA CTC CTG CTT GAA GAC GAG GAT GCA GCC GCG GCT GGA GGC GGG GGT GTA GTC GTG GTT TAA TAC TAG TAT TCA TCC TCG TCT TGA TGC TGG TGT TTA TTC TTG TTT
Ala 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
Arg 0 0 0 0 0 0 0 0 1 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
Asn 0 1 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
Asp 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
Cys 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
Gln 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
Glu 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
Gly 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
His 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
Ile 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
Leu 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 1 0
Lys 1 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
Met 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
Phe 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 1 0
Pro 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
Ser 0 0 0 0 0 0 0 0 0 1 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
Thr 0 0 0 0 1 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
Trp 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
Tyr 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
Val 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
Stop 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0

(8)

One can define three different channels for this problem. The codon-codon channel, the codon-aminoacid channel and the aminoacid-aminoacid channel. In [16] and [17], Bouaynaya *et al.*, study the information transfer process between DNA and aminoacids, underlining the breakdown of the communications system analogy and propose modelling the process with an aminoacid channel. That is, both the transmitted (X) and received (Y) signals are aminoacids assuming a virtual protein source to DNA encoder. They characterised the communication channel using first the PAM250 matrix due to Dayhoff et al. [18] and then by an aminoacid 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 aminoacids and therefore the codon substitution matrix needs to be propagated over the generations, and not the aminoacid substitution matrix. However, the "meaning" of the message is in aminoacids.

Using the Kimura nucleotide substitution model, we generate the corresponding codon (three-nucleotide) 64×64. We propagate the message in the form of codons over generations in each of the three channels and then encode the received codon to aminoacid and calculate the capacity based on this channel and encoder.

5. Optimality of the natural code

It is curious that the natural genetic code (mapping) is not uniform. While some aminoacids are coded by 6 different codons, some are coded by

4, 3, 2 or 1 codons. A natural question to ask is whether the natural genetic code is optimal in the information preservation, or channel capacity sense. We have constructed a number of alternatives to the natural code:

1. a degenerate code where all the aminoacids are coded by only 1 codon and the remaining 24 codons are the stop codons.
2. a uniform code in which all aminoacids are coded by 3 codons (and the stop codon by $64 - 20 \times 3 = 4$) which we will call the uniform3-code.
3. an almost uniform code in which the aminoacids are coded by 4 or 2 codons, which we will call the uniform $4-2$-code.
4. a code obtained from the natural code by flipping C and G and A and T, that is transitions and transversions are interchanged which we will call the flipped natural code.
5. similarly flipped version of the uniform $4-2$ code.

We have calculated the channel capacities for these alternative codes using the Blahut-Arimoto algorithm, which are presented in Figure 9. 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 or the degenerate code has the lowest channel capacity irrespective of the value of K. The flipped natural code has higher channel capacity when $K > 1$, in which case transversions rather than transitions are favoured. The uniform-3 code has one of the lower channel capacity curves and takes over 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, fall slightly behind a uniform $4-2$ code. This final result is important to state that natural code is not neces-

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

Figure 6: The degenerate (extreme) genetic code (codon to aminoacid 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

Figure 7: The uniform-3 genetic code (codon to aminoacid 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

Figure 8: The uniform-42 genetic code (codon to aminoacid 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

sarily the optimal code at least in terms of channel capacity or information preservation or robustness to mutations.

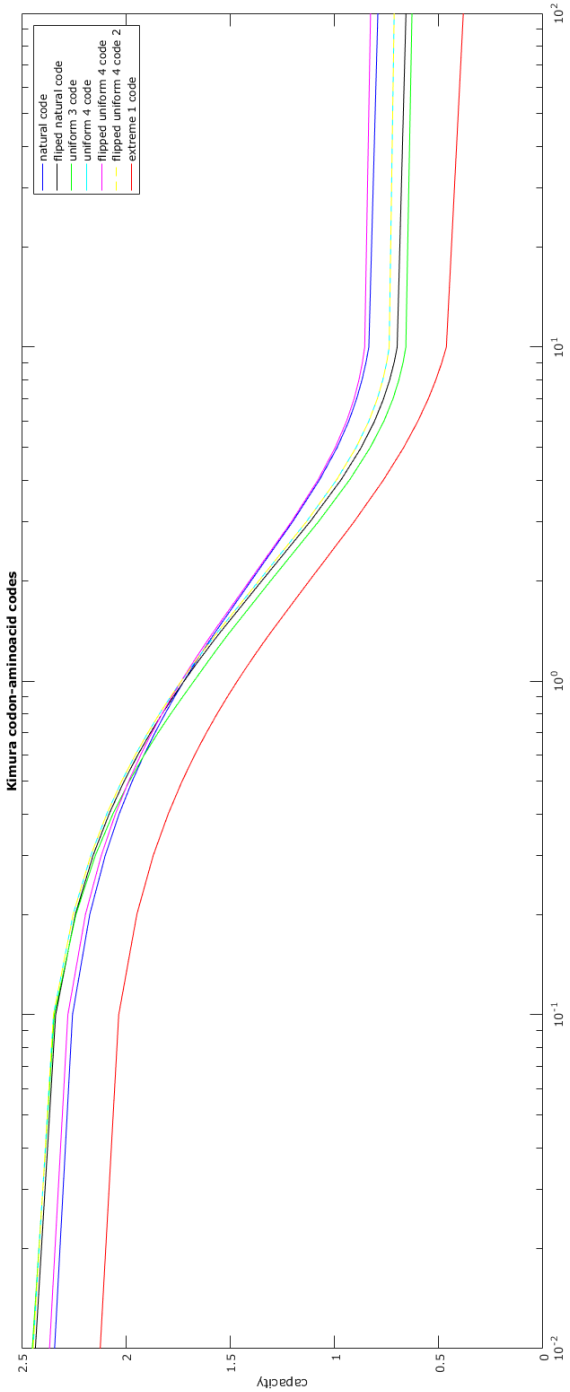


Figure 9: Comparison of various synthetic genetic codes and the natural genetic code.

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 F. Crick who proposed the "frozen accident" model [19]. The "frozen accident" model was questioned by various researchers in the literature who noted the "superiority" of the natural code to alternatives. For example, Hurst and Freeland and Hurst [20] generated $1,000,000$ 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". Various 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. [21]).

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 second order statistics. The use of MSE measure intrinsically makes a Gaussian distribution assumption which is not necessarily suggested by the nature of the data. We also do not use hydrophobicity but the Hamming distance between the codes since we think that the unequal transition rates for transitions and transversions modelled in the substitution matrix already takes care of the physical facts. The searches made in the literature seem to be random picks of codes from the space of possible codes such as in [20] which generated $1,000,000$ different configurations but as noted in [22], 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 on. We propose an intelligent search algorithm which learns through its search and searches at increasingly more probable parts of the space for solutions.

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" (aminoacids), unlabelled since we can rename the aminoacids 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 $S(64,21) = 2.9\times10^{29}$. 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, by remapping a codon to a new aminoacid and calculate the channel capacity for all such generated new configurations. There are $64 \times 20 = 1280$ such configurations (2 Hamming distance neighbours of the natural code). No configurations at Hamming distance 2 gave a higher channel capacity then the natural code. Below in Figure 10, we provide the histogram of the capacities of all such configurations:

This result shows us that the natural code is at least at a local optimal. However, this result does not generalise to neighbours at greater Hamming distances than 2. We can construct a higher capacity code at 4-Hamming distance 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 the natural code, we see that the codons are mostly coded in groups of 4

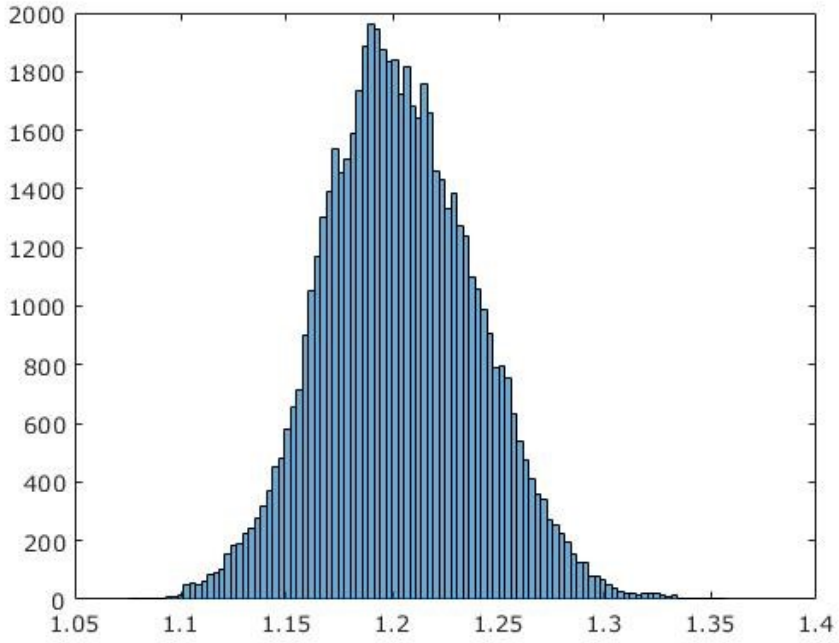


Figure 10: Histogram of the genetic codes 2 Hamming distance from the natural code.

or 2 to an aminoacid 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 neighbour code to the natural code by moving TGA from STOP to Tryptophan and ATA from Isoleucine to Methionine as depicted in Figure 11. The resulting code is at 4-Hamming distance from the natural code. As can be seen in Figure 12, the channel capacity curve of this code is slightly above that of the natural code. Therefore, one can conclude that the optimality of the natural code is very local and does even extend to a neighbourhood

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

Figure 11: A genetic code 4-Hamming distance from the natural code (codon to aminoacid 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

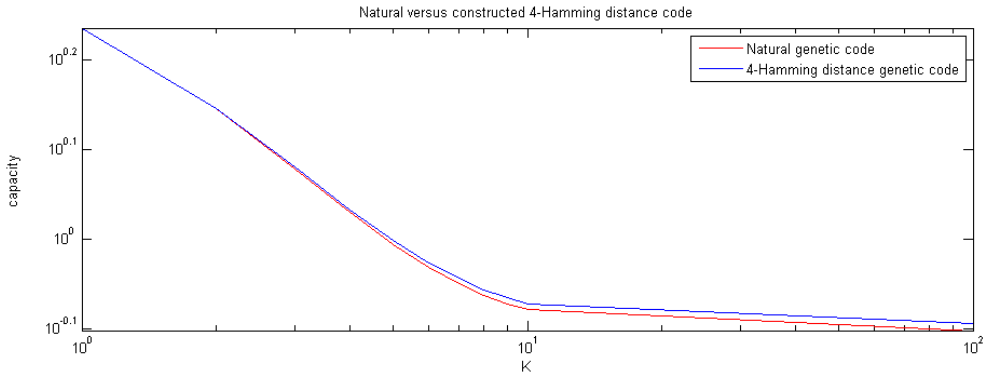


Figure 12: Comparison of channel capacities for the natural genetic code and a constructed genetic code 4-Hamming distance from the natural code.

4-Hamming distance. This observation motivates one to look for a globally optimal codon-aminoacid code.

As mentioned above although several attempts exist in the past to search for an optimal code, only random non exhaustive searches were made covering far less than a statistically meaningful space. The searches were not intelligent leading to non-conclusive results. To search for a global optimal, we propose to use a non-convex optimization algorithm, namely Simulated Annealing algorithm to do an intelligent search of the best code. Simulated Annealing algorithm has had success in a wide variety application areas where the optimization problem at hand was 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.

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 [23]. Simulated Annealing proceeds with series of random walks, namely Metropolis loops during which new configurations are proposed. If the new configuration leads to a better cost (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 problem 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 are accepted. It has been proved that if a logarithmic cooling schedule is applied the algorithm converges to the global optimal. However, logarithmic cooling scheme can get infinitely slow and suboptimal schemes such as geometric cooling scheme is applied. For detailed information on the simulated annealing algorithm, one is referred to [24]. 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 $T > T_{min}$
  - $T \leftarrow T \times \alpha$
  - 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, move to the new state
  - $M \leftarrow M_{new}$

- Output: the final code M .

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 one 1 to a 0in the aminoacid-codon matrix. That is, changing the mapping of one codon from one aminoacid to another aminoacid making sure that there is at least one codon assigned to each aminoacid.

Figure 13 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 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 average improvement is fast, reducing slowly and then saturating to significantly better codes or high capacity with small oscillations around the "near-optimal" codes.

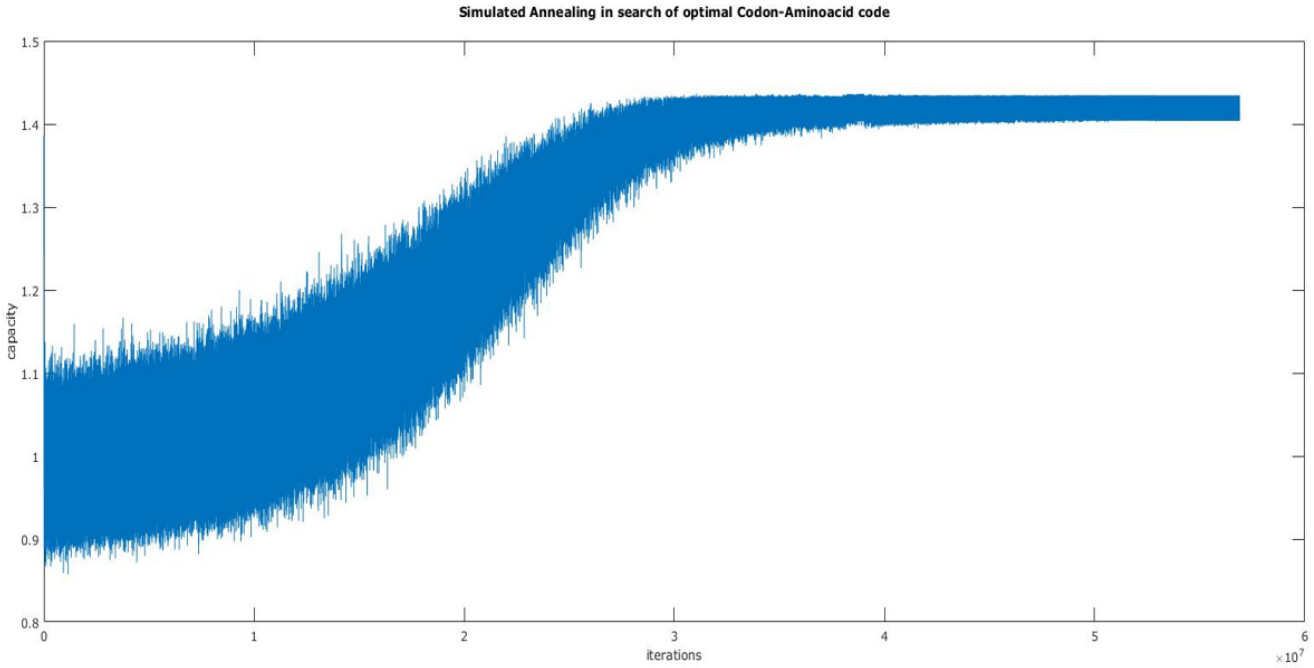


Figure 13: Channel capacity for Kimura codon-aminoacid channel for various values of mutation rate.

The best such found configuration is given in Figure 14. It is very interesting to note that as in the case of the natural code, the codons producing the same aminoacid are close in the table and have ambiguities in the nucleotides. The ambiguities in this optimal code are in the first (10), 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 capacity profiles of this near optimal code with the natural code in Figure 15. To give a scale of comparison, the capacity curves of the degenerate code (one codon synthesizing one aminoacid) 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 found has significantly higher information capacity then 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.

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

Figure 14: The uniform-42 genetic code (codon to aminoacid 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

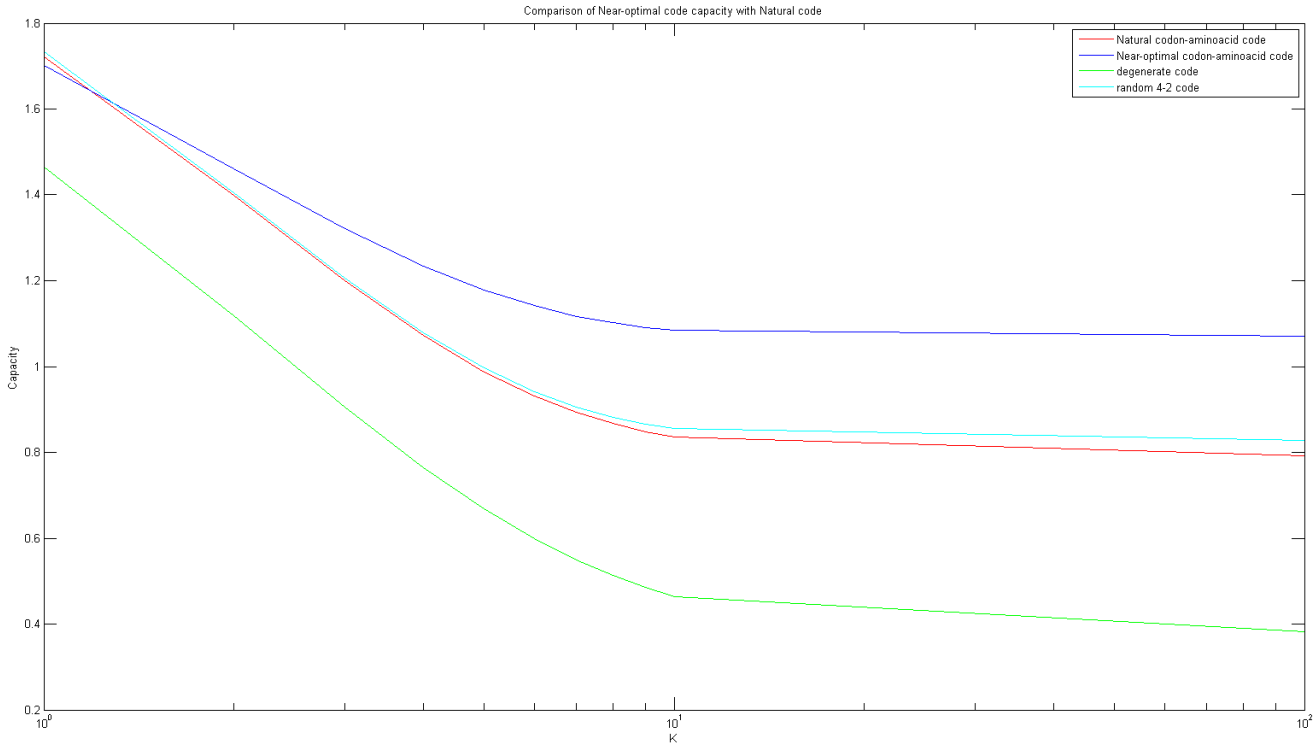


Figure 15: Comparison of Channel capacity of Natural, Near-Optimal, Degenerate, Random 4-2 codes on Kimura codon-aminoacid channel for various values of mutation rate at N generations.

These observations need a discussion on the biological significance. In particular, they underline clearly that the natural codon-aminoacid code/map is far from being optimal. The natural code can be "one in a million" []; however, considering that there are more than 10^{29} possible configurations, this is not a statistically significant measure. There are many other codes that have far better information preservation capabilities.

This observation may indirectly give support to two hypothesis.

1. frozen accident hypothesis of Crick [25].
2. that 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 optimised for 3-codons and that almost all redundancies are in the third position.

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.

6. 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-aminoacid channel rather than aminoacid-aminoacid 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 aminoacid) 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 significantly extended our search space (close to 60 million tested configurations) but more importantly we have done our search not "blindly" but "intelligently" using a non-convex learning/optimisation algorithm, namely Simulated Annealing. Our observations provide strong evidence for the theory that once the codons were formed of 2-codons only and that the third nucleotide was acquired later.

7. Acknowledgements

This project was principally funded by the Alexander von Humboldt Foundation in the form of an Experienced Research Fellowship awarded to EE Kuruoglu. The authors would like to thank Prof Martin Vingron and Prof Alexander Bolshoy whose comments helped improve this work.

References

[1] T. Jukes, L. Gatlin, Recent studies concerning the coding mechanism., Progress in nucleic acid research and molecular biology 11 (1971) 303–350.

[2] H. Yockey, Can the central dogma be derived from information theory?, Journal of Theoretical Biology 74 (1) (1978) 149–152.

[3] R. Román-Roldán, P. Bernaola-Galván, J. Oliver, Application of information theory to dna sequence analysis: A review, Pattern Recognition 29 (7) (1996) 1187–1194.

[4] C. Shannon, A mathematical theory of communication, Bell System Technical Journal 27 (3) (1948) 379–423.

[5] T. Schneider, J. Spouge, Information content of individual genetic sequences, Journal of Theoretical Biology 189 (4) (1997) 427–441.

[6] T. Schneider, Evolution of biological information, Nucleic Acids Research 28 (14) (2000) 2794–2799.

[7] T. Schneider, A brief review of molecular information theory, Nano Communication Networks 1 (3) (2010) 173–180.

[8] T. Cover, J. Thomas, Elements of Information Theory, Wiley, 2005.

[9] R. Blahut, Computation of channel capacity and rate-distortion functions, IEEE Transactions on Information Theory 18 (4) (1972) 460–473.

[10] S. Arimoto, An algorithm for computing the capacity of arbitrary discrete memoryless channels, IEEE Transactions on Information Theory 18 (1) (1972) 14–20.

[11] T. Jukes, Recent problems in the genetic code., Current Topics in Microbiology and Immunology 49 (1969) 178–219.

[12] M. Nei, S. Kumar, Molecular Evolution and Phylogenetics, Oxford University Press, 2000.

[13] C. Ponting, P. Oliver, W. Reik, Evolution and functions of long noncoding {RNAs}, Cell 136 (4) (2009) 629 – 641.

[14] M. Kimura, A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences, Journal of Molecular Evolution 16 (2) (1980) 111–120.

[15] J. Felsenstein, Evolutionary trees from dna sequences: A maximum likelihood approach, Journal of Molecular Evolution 17 (6) (1981) 368–376.

[16] N. Bouaynaya, D. Schonfeld, Protein communication system: Evolution and genomic structure, Algorithmica (New York) 48 (4) (2007) 375–397.

[17] L. Gong, N. Bouaynaya, D. Schonfeld, Information-theoretic model of evolution over protein communication channel, IEEE/ACM Transactions on Computational Biology and Bioinformatics 8 (1) (2011) 143–151.

[18] B. C. Dayhoff M. O.; Schwartz, R. M.; Orcutt, A model of evolutionary change in proteins, Atlas of Protein Sequence and Structure 5 (3) (1978) 345352.

[19] F. Crick, The origin of the genetic code, Journal of Molecular Biology 38 (3) (1968) 367–379.

[20] S. Freeland, L. Hurst, The genetic code is one in a million, Journal of Molecular Evolution 47 (3) (1998) 238–248.

[21] S. Freeland, T. Wu, N. Keulmann, The case for an error minimizing standard genetic code, Origins of Life and Evolution of the Biosphere 33 (4-5) (2003) 457–477.

[22] J. Santos, . Monteagudo, Genetic code optimality studied by means of simulated evolution and within the coevolution theory of the canonical code organization, Natural Computing 8 (4) (2009) 719–738.

[23] S. Kirkpatrick, C. D. Gelatt, M. P. Vecchi, Optimization by simulated annealing, SCIENCE 220 (4598) (1983) 671–680.

[24] P. van Laarhoven, E. Aarts, Simulated Annealing: Theory and Applications, Springer, 1987.

[25] F. Crick, Codon-anticodon pairing: The wobble hypothesis, J. Mol. Biol (1966) 548–555.