

## Rapporto Finale: Short Term Mobility

Ercan Engin Kuruoglu, ISTI-CNR, Pisa

presso Max Planck Institute for Molecular Genetics

Durante il periodo di Short Term Mobility l'attività di ricerca è stata svolta sugli argomenti seguenti:

1. Modellare processi di sintesi di proteine tramite sistemi di telecomunicazione
2. Ricerca di codifiche DNA-proteine ottimali dal punto di vista di conservazione dell'informazione
3. Modellare interazioni genomiche che cambiano nel tempo tramite reti Bayesiane

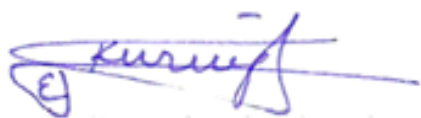
I risultati ottenuti sugli argomenti 1 e 2 sono stati riportati nell'articolo allegato dove è stata descritta in dettaglio la nostra ricerca. Questo articolo è stato sottomesso alla rivista *Journal of Theoretical Biology*. In breve:

- Abbiamo sviluppato un modello matematico per rappresentare la sintesi di proteine che utilizza le teorie di Shannon.
- Abbiamo calcolato i limiti di conservazione dell'informazione (channel capacity) sotto diversi modelli statistici per le mutazioni
- Abbiamo mostrato l'andamento della perdita dell'informazione durante successive generazioni
- Abbiamo sviluppato un algoritmo "intelligente" per cercare la codifica ottimale DNA-proteina e mostrato che la codifica naturale è "sotto-ottimale". Questa osservazione dà supporto alla teoria di "frozen accident" di Francis Crick.
- È stato osservato inoltre che la codifica naturale ha delle ambiguità solo nella terza posizione dei codoni, mentre la codifica ottimale ha delle ambiguità in tutte e tre le posizioni dei codoni. Questo dimostra che i codoni originali erano fatti da 2 nucleotidi e ad un certo punto, durante l'evoluzione, è stato aggiunto il terzo nucleotide.

Inoltre, durante questa attività di ricerca, abbiamo fatto riunioni con altri membri dell'istituto, incluso il direttore Prof. Martin Vingron, su modelli delle interazioni genomiche che cambiano nel tempo e abbiamo identificato un nuovo approccio per le nostre future ricerche, sempre basato sul nostro lavoro già pubblicato recentemente: "Time varying gene network modelling using sequential Monte Carlo," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, Novembre 2016.

Abbiamo fatto anche riunioni con un gruppo di ricerca presso la Free University (Computational Proteomics di Prof Tim Conrad) di Berlino sulla implementazione di questo algoritmo su un loro super computer. Abbiamo deciso di preparare una proposta di progetto per Horizon2020 per finanziare questa attività di ricerca.

Altri incontri sono stati fatti sulle "conformation maps" del DNA con Prof Martin Vingron ed è stato deciso di esplorare questa area per una collaborazione futura.



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

Ercan E. 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. We calculate Information capacities of DNA on the nucleotide (for non-coding DNA) and aminoacid (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-aminoacid 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 aminoacids.

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

---

<sup>1</sup>Corresponding author

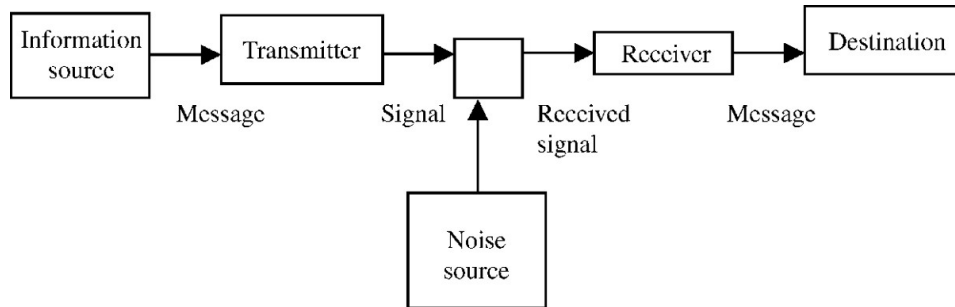


Figure 1: A generic communications system.

## 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  
 5 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  
 10 an aminoacid 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 (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 material and energy, and robust to  
 15 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 [1, 2, 3, 4, 5]. A key element of the analogy is the ability to quantify the information which is provided by the *entropy* as an information measure [6]. Numerous publications in the literature have studied the entropy of the DNA [7], across the species, at protein binding sites [8, 9], etc. The reader is referred to the paper by Fabris [10] 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 [11, 12]. Only a few works [13, 14], 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-aminoacid) 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

45 must be underlined that a full analogy with a communication system fails in the sense the encoder is lacking in a biological system. In the case of protein coding, the decoded message is not a DNA but an aminoacid sequence. In this case, one can at best talk of a hypothetical information source already coded in the form of a nucleotide sequence.

50 In this article, utilizing the Coding Theory of Shannon, we develop theoretical limits of information preservation in non-coding and aminoacid 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  
55 of the optimality of the natural genetic (codon-aminoacid) 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-aminoacid codes were explored. With this publication, we propose an "intelligent" search  
60 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  
65 coding theory and define channel capacity. We give channel capacity results on non-coding DNA and protein coding DNA in Section 2.2 and Section 2.3, respectively. The optimality of the natural codon-aminoacid encoder is studied in Section 3. Conclusions and future research directions are provided in Section 4.

## 70 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 [6]:

$$H(p) = - \sum_i p_i \log_2 p_i, \quad (1)$$

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  $p_{[A,C,G,T]} = [0.29 \ 0.21 \ 0.21 \ 0.29]$ , the entropy is calculated to be  
75  $H(p_{[A,C,G,T]}) = 1.9815 < 2$ . 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  
 $H(p_{codons}) = 5.7936 < 3 \times H(p_{[A,C,G,T]}) = 5.9445$ . If all the codons were equiprobably distributed it would have achieved the maximum value of 6.  
80 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 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.

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

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  
85 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 [15]. However, a numerical algorithm exists for calculating the channel capacity in the other cases, which is called the Blahut-Arimoto algorithm [16, 17]. 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 characterized 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 aminoacids, 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 [18]. Hence, the substitution matrix is characterized with only one parameter, the

nucleotide substitution rate  $q$ . The Jukes-Cantor rate matrix is given in

$$Q_{JC} = \begin{bmatrix} -3q & q & q & q \\ q & -3q & q & q \\ q & q & -3q & q \\ q & q & q & -3q \end{bmatrix} \quad (3)$$

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 ([19])

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

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  
105 channels in Figure 1 is

$$C_m = \max_p I(X; Y(m)) = \max_p [H(Y(m)) - H(Y(m)|X)] \quad (5)$$

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

The results are given in Figure 2 which show the exponential decline of information capacity of the non-coding DNA channel with increasing num-  
115 ber of generations. The results show clearly that information (capacity)



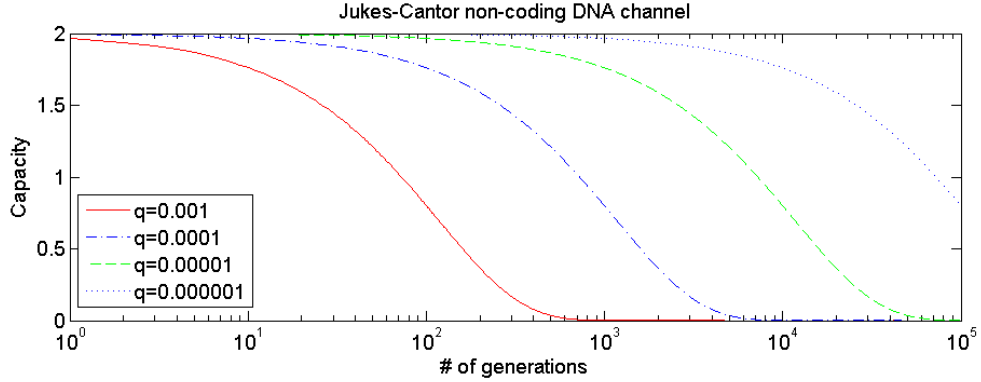


Figure 2: Channel capacity for Jukes-Cantor non-coding DNA channel for various values of mutation rate in units of generation length.

vanishes exponentially over generations and that the time scale is given by the mutation rate.

In the biological context, the rates of substitutions rates for 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 model, which takes care of this effect, exists due to Kimura [20]. 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

120  $C_m = \max_p I(X; Y(m)) = \max_p [H(Y(m)) - H(Y(m)|X)]$ . The capacity

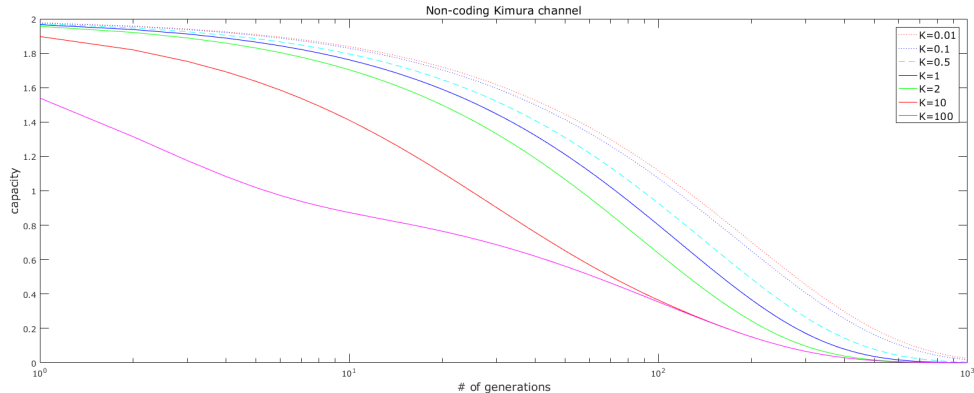


Figure 3: Channel capacity for Kimura non-coding DNA channel for various values of transitions/transversions rate ratio  $K$ .  $q = 0.001$ .

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 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 [21]. The Felsenstein substitution rate matrix is given by:

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

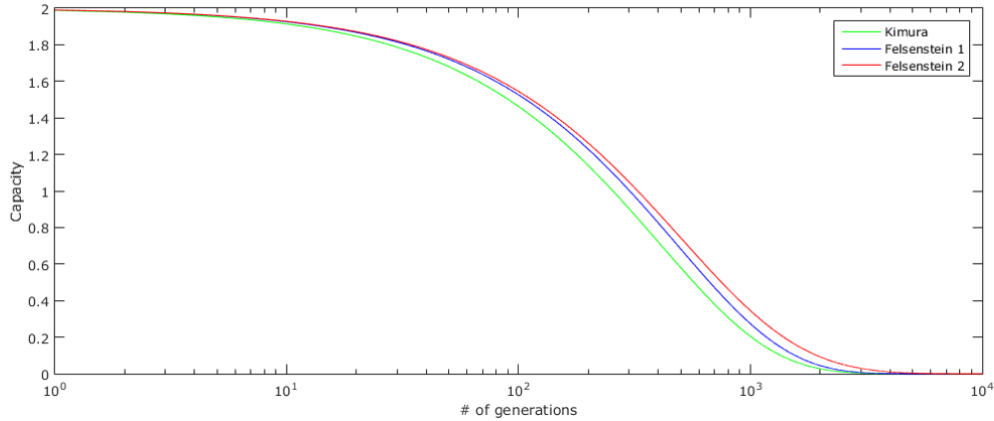


Figure 4: Channel capacity for Felsenstein non-coding DNA channel for two values of mutation rate and comparison with Kimura channel. Kimura parameter:  $K = 2$ , Felsenstein parameters:  $\pi_1 = [0.3 \ 0.2 \ 0.2 \ 0.3]$ ,  $\pi_2 = [0.4 \ 0.2 \ 0.3 \ 0.1]$

130 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  
 135 substitution 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.

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  
 140 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 [22].

The channel capacity of non-coding DNA can provide us an intuition to  
145 what extend these functions can be preserved. It must be noted, however,  
that unlike the 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 depen-  
150 dent 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  
155 output aminoacids, 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  
aminoacids and some are used as stop markers. There is redundancy in the  
codon-aminoacid mapping and this redundancy is used as an error correcting  
mechanism. The mapping between codons and aminoacids is given in Figure  
160 5. This mapping can also be represented in matrix form as in Eq. (8).

1	T	C	A	G	3
2					
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 the 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.



One can define three different channels for this problem. The codon-codon channel, the codon-aminoacid channel and the aminoacid-aminoacid channel. In [23] and [13], Bouyanaya *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  
165 an aminoacid-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.* [24] and then by an aminoacid  
170 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  
175 therefore the codon substitution matrix needs to be propagated over generations, and not the aminoacid substitution matrix. However, one should keep in mind that the "meaning" of the message is in aminoacids.

Using the Kimura nucleotide substitution model, we generate the corresponding codon (three-nucleotide)  $64 \times 64$  substitution matrix. We propagate the message in the form of codons over generations and then decode  
180 the received codon to an aminoacid 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.  
185 While most of the aminoacids are coded by 2 different codons, some are

coded by 6, 4, 3 or 1 codons. A natural question to ask is whether the natural genetic code is optimal in the information preservation, or channel capacity sense. To have an understanding of the space of possible codon-aminoacid mappings, we have constructed a number of alternatives to the  
190 natural code:

1. an extreme-1 code where each aminoacid is coded by only 1 codon and the remaining 44 codons are stop codons (Figure 6).
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 uniform 3  
195 code (Figure 7).
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 (Figure 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 aminoacid for 2-fold degenerate codons. We will call this the flipped  
200 natural code (Figure 9).
5. similarly flipped version of the uniform 4-2 code (Figure 10).

We have calculated the channel capacities for the natural aminoacid code as well as the alternative codes using the Blahut-Arimoto algorithm, which  
205 are presented in Figure 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  
210 capacity when  $K < 1$ , in which case transversions rather than transitions on the 3rd codon do not change the aminoacid for 2 fold degenerate codons.



1	T	C	A	G	3
2					
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	T	C	A	G	3
2					
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	T	C	A	G	3
2					
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

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

Figure 9: 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

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

Figure 10: 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

The uniform-3 code has one of the lower channel capacity curves and surpasses 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  
215 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.

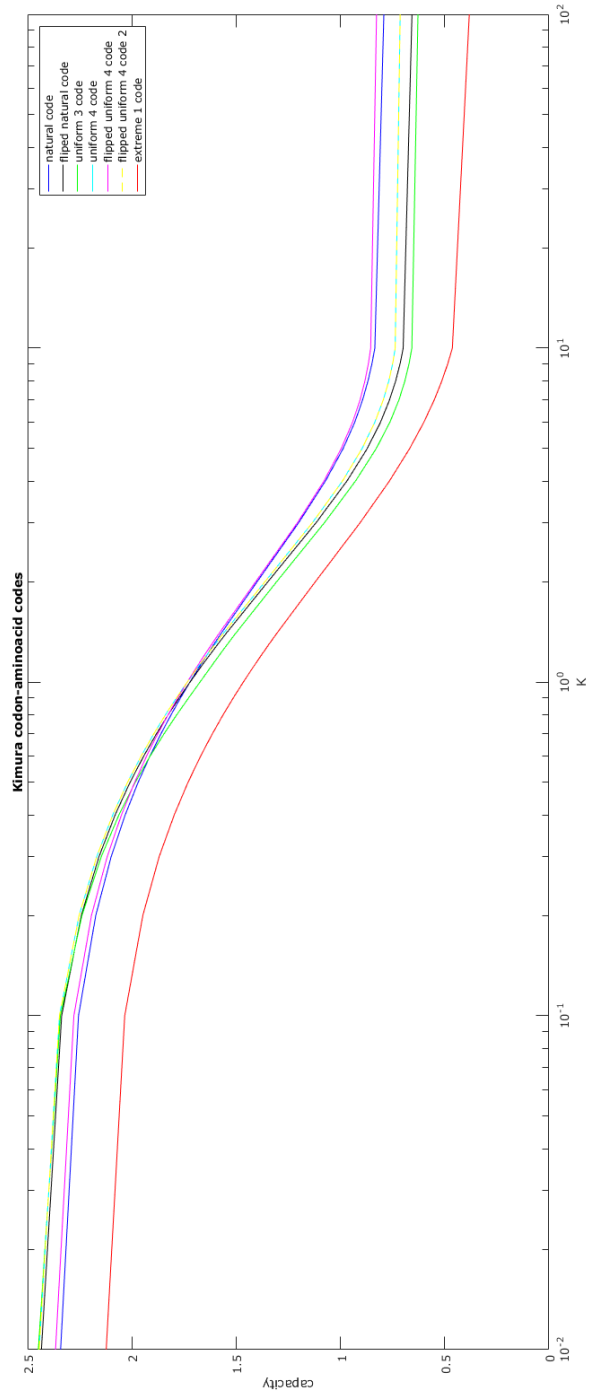


Figure 11: Comparison of the Kimura channel capacities versus  $K$  ( $q = 0.001$ ) for various synthetic genetic codes and the natural genetic code. The values at 100th generation are plotted.

These observations make us ask the question why the natural code was  
220 preferred to any other code. This question was asked before by several re-  
searchers including Crick who proposed the "frozen accident" model [25].  
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 [26] generated randomly 1,000,000 dif-  
225 ferent 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".

Most researchers use the *polar requirement*, a measure of hydrophobicity  
as the error measure and try to find/produce codes that minimize this cost  
230 function (e.g. [27]). The reader is referred to also an interesting review by  
Tlustý [28]. Our approach is different from previous work in a number of  
aspects. Rather than using MSE (mean square error) on specific biochem-  
ical properties such as hydrophobicity, we use an information theory based  
measure which captures information on all statistics rather than only the  
235 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 [26] which gen-  
erated 1,000,000 different configurations but as noted in [29], the explored  
240 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 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  
245 reported in [29], however, they utilize a genetic algorithm rather than sim-



ulated 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 [30] 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" (aminoacids), unlabelled since we can rename the aminoacids without losing any biological meaning. This is 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:

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

250 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  $1.23 \times 10^{63}$ . 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  
 255 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 aminoacid and calculate the channel capacity for all such generated new configurations. While doing this we ensure that  
 260 all aminoacids are encoded by at least one codon. Disregarding the case of rows with a single 1,  $62 \times 20 = 1240$  such configurations (Hamming distance 2 neighbours of the natural code). Below in Figure 12, we provide the histogram of the capacities of all such configurations: The natural code is

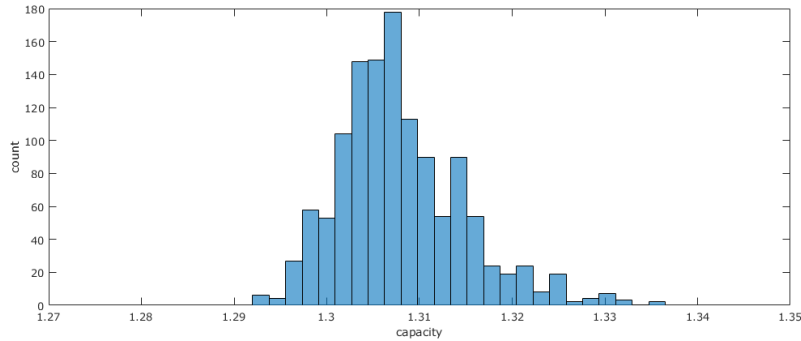


Figure 12: Histogram of capacities (Kimura model) the genetic codes at Hamming distance 2 from the natural code. The natural code has capacity 1.3219.  $K = 2, q = 0.001$ .

one of the best but not the best among its neighbours in terms of capacity.

265 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 the natural code, we see that the codons are mostly coded in groups of 4 or 2 to an aminoacid with redundancies mostly at the third codon position and less at the first codon

270 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 Figure 13. The resulting code is at

275 Hamming distance 4 from the natural code. As can be seen in Figure 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-aminoacid maps in the vicinity of the natural code. Either nature did not care to optimize

2	1	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 13: 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

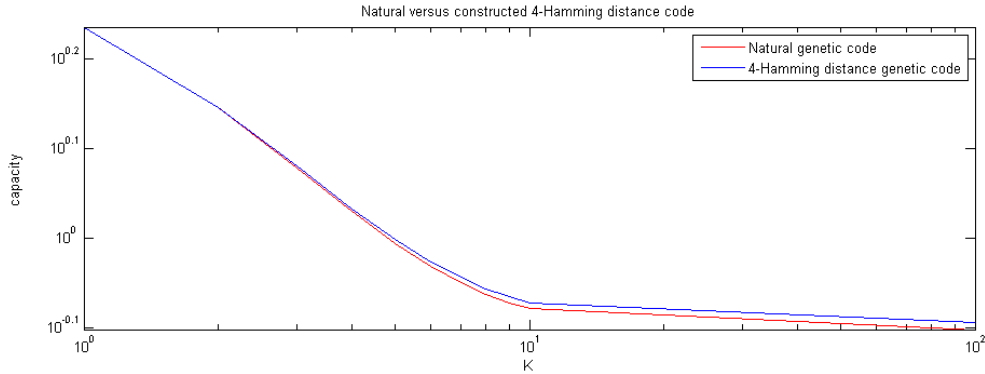


Figure 14: Comparison of channel capacities for the natural genetic code and a constructed genetic code at Hamming distance 4 from the natural code. Kimura channel ( $K=2, q=0.001$ ), 100 generations.

280 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  
 285 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  
 290 results. To search for a global optimum, we propose to use a non-convex optimization algorithm, namely *Simulated Annealing* algorithm [31], to do an intelligent search of the optimal code. The Simulated Annealing algorithm has had success in a wide variety application areas where the optimization problem at hand is NP-hard, that is not solvable in polynomial time. These  
 295 application areas include the traveling salesman problem, graph partition-

ing, scheduling in operations research, VLSI circuit design in electronics, optimal source coder design in telecommunications, etc [31, 32].

Simulated Annealing is motivated by experimental solid state physics where solids are first heated to a very high temperature and then cooled  
300 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 [31]. Simulated Annealing proceeds with a series of random walks, namely Metropolis loops during which new configurations are proposed. If the new configuration  
305 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  
310 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  
315 as a geometric cooling scheme is applied. For detailed information on the simulated annealing algorithm, one is referred to [33]. A brief sketch of the algorithm is given below:

### 320 **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  $N(\cdot)$  includes all 2-Hamming distance codes from the code  $M$
  - If  $P(C(M), C(M_{new}), T) \geq \text{random}(0, 1)$ , where  $C(\cdot)$  is the channel capacity and  $P(\cdot)$  is the Boltzmann function,
    - \* then move to the new state  $M \leftarrow M_{new}$
- Output: the final code  $M$  .

330 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 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. 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.

345 Figure 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 re-  
350 cover 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  
355 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  
360 with a rapid increase in the capacity values as expected since the extreme code is a degenerate code with only one codon mapping to each aminoacid.

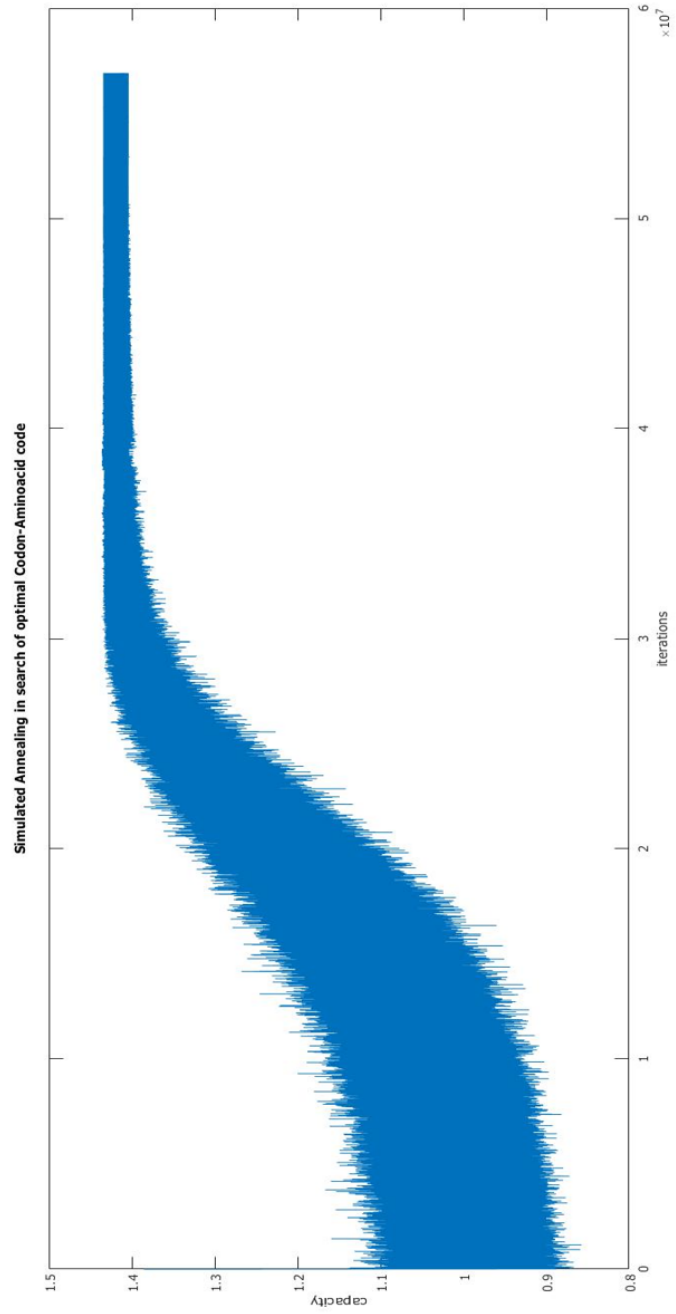


Figure 15: The capacity of the code during the evolution of Simulated Annealing algorithm. The capacity value of the natural code is 1.38. Initial temperature is 0.1, temperature coefficient is 0.99.



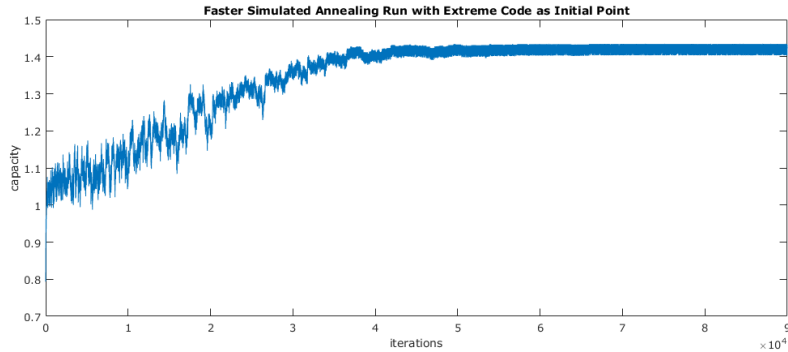


Figure 16: The capacity of the code during the evolution of a fast run of Simulated Annealing algorithm with the extreme code as the initial code. The capacity value of the extreme code is 0.79. Initial temperature is 0.01, temperature coefficient is 0.95.

The best configuration found in the simulations is given in Figure 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  
 365 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 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  
 370 (0) position.

We provide a comparison of capacity profiles of this near optimal code with the natural code in Figure 18. 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  
 375 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

1	T	C	A	G	3
2					
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 17: 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

obtained by the Simulated Annealing algorithm has significantly higher information capacity than the natural code. The difference is at the same  
380 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.

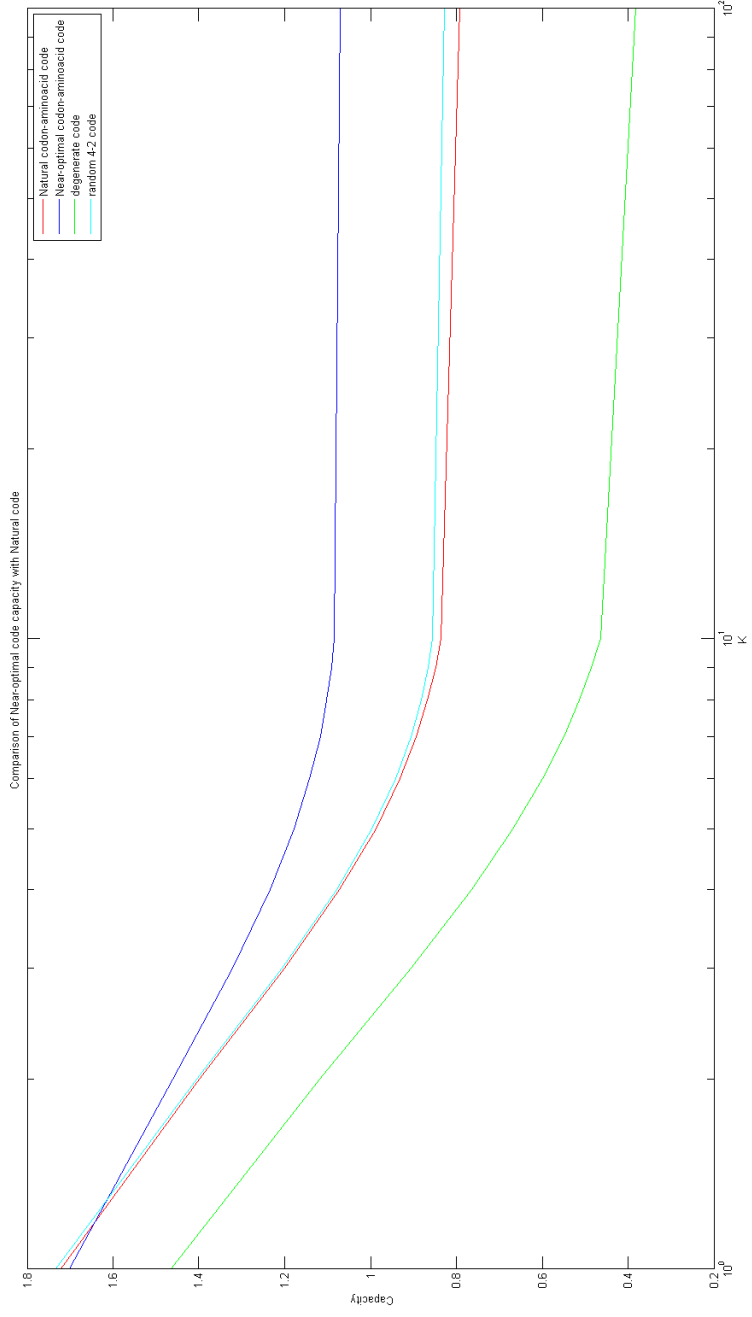


Figure 18: Comparison of Channel capacity of Natural, Near-Optimal, Degenerate, Random 4-2 codes on Kimura codon-aminoc acid channel for various values of mutation rate at  $N$  generations.

385        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 although being better than most possible codes.  
The natural code can be "one in a million" [26]; however, considering that  
there are more than  $10^{63}$  possible configurations, being one in a million is  
390 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 [25, 30], so its evolution was stopped  
395 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 [34]
3. that at some point in the past the codons were composed of 2 nu-  
cleotides only and the third nucleotide was acquired afterwards. This  
400 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  
[35, 29].

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  
405 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.

## 4. Conclusions

410 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  
415 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 demon-  
strate the suboptimality of the natural code without any space for doubt.  
Its channel capacity is significantly below that of various other codes. Un-  
420 like 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  
425 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  
430 later. The presented formulation, which places the information capacity as  
a measure of 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  
435 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

- 440 [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.
- 445 [3] R. Román-Roldán, P. Bernaola-Galván, J. Oliver, Application of infor-  
mation theory to dna sequence analysis: A review, Pattern Recognition  
29 (7) (1996) 1187–1194.
- [4] G. Battail, An engineers view on genetic information and biological  
evolution, Biosystems 76 (13) (2004) 279–290.
- 450 [5] A. Konopka, Information theories in molecular biology and genomics,  
eLS.
- [6] C. Shannon, A mathematical theory of communication, Bell System  
Technical Journal 27 (3) (1948) 379–423.
- [7] T. Schneider, J. Spouge, Information content of individual genetic se-  
455 quences, Journal of Theoretical Biology 189 (4) (1997) 427–441.

- [8] T. Schneider, Evolution of biological information, *Nucleic Acids Research* 28 (14) (2000) 2794–2799.
- [9] T. Schneider, A brief review of molecular information theory, *Nano Communication Networks* 1 (3) (2010) 173–180.
- 460 [10] F. Fabris, Shannon information theory and molecular biology, *Journal of Interdisciplinary Mathematics* 12 (1) (2009) 41–87.
- [11] E. May, M. Vouk, D. Bitzer, D. Rosnick, An error-correcting code framework for genetic sequence analysis, *Journal of the Franklin Institute* 341 (1-2) (2004) 89–109.
- 465 [12] G. Battail, Can we explain the faithful communication of genetic information?, in: P. Siegel, S. E, V. AJ, V. B (Eds.), *Advances in Information Recording, Vol. 8*, American Mathematical Society: DIMACS-Series in Discrete Mathematics and Theoretical Computer Science, 2004, Ch. 10, pp. 79–103.
- 470 [13] 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.
- [14] F. Balado, Capacity of dna data embedding under substitution mutations, *IEEE Transactions on Information Theory* 59 (2) (2013) 928–941.
- 475 [15] T. Cover, J. Thomas, *Elements of Information Theory*, Wiley, 2005.
- [16] R. Blahut, Computation of channel capacity and rate-distortion functions, *IEEE Transactions on Information Theory* 18 (4) (1972) 460–473.



- [17] S. Arimoto, An algorithm for computing the capacity of arbitrary discrete memoryless channels, *IEEE Transactions on Information Theory* 18 (1) (1972) 14–20.  
480
- [18] T. Jukes, Recent problems in the genetic code., *Current Topics in Microbiology and Immunology* 49 (1969) 178–219.
- [19] M. Nei, S. Kumar, *Molecular Evolution and Phylogenetics*, Oxford University Press, 2000.  
485
- [20] 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.
- [21] J. Felsenstein, Evolutionary trees from dna sequences: A maximum likelihood approach, *Journal of Molecular Evolution* 17 (6) (1981) 368–376.  
490
- [22] C. Ponting, P. Oliver, W. Reik, Evolution and functions of long non-coding {RNAs}, *Cell* 136 (4) (2009) 629 – 641.
- [23] N. Bouaynaya, D. Schonfeld, Protein communication system: Evolution and genomic structure, *Algorithmica (New York)* 48 (4) (2007) 375–397.  
495
- [24] 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.
- [25] F. Crick, The origin of the genetic code, *Journal of Molecular Biology* 38 (3) (1968) 367–379.  
500

- [26] S. Freeland, L. Hurst, The genetic code is one in a million, *Journal of Molecular Evolution* 47 (3) (1998) 238–248.
- [27] 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.  
505
- [28] T. Tlusty, A colorful origin for the genetic code: Information theory, statistical mechanics and the emergence of molecular codes, *Physics of Life Reviews* 7 (3) (2010) 362376.
- [29] 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.  
510
- [30] G. Sella, D. Ardell, The coevolution of genes and genetic codes: Crick’s frozen accident revisited, *J Molecular Evolution* 63 (3) (2006) 297–313.
- [31] S. Kirkpatrick, C. D. Gelatt, M. P. Vecchi, Optimization by simulated annealing, *Science* 220 (4598) (1983) 671–680.  
515
- [32] E. Kuruoglu, E. Ayanoglu, The design of finite-state machines for quantization using simulated annealing, in: R. Calderbank, G. Forney, N. Moayeri (Eds.), *Coding and Quantization: DIMACS/IEEE Workshop*, October 19-21, 1992, American Mathematical Society: DIMACS-Series in Discrete Mathematics and Theoretical Computer Science, 1993, pp. 175–184.  
520
- [33] P. van Laarhoven, E. Aarts, *Simulated Annealing: Theory and Applications*, Springer, 1987.

- 525 [34] D. Ardell, G. Sella, No accident: genetic codes freeze in error-correcting patterns of the standard genetic code, *Philosophical Transactions of the Royal Society B: Biological Sciences* 357 (1427) (2002) 1625–1642.
- [35] G. Sella, D. Ardell, Possibilities for the evolution of the genetic code from a preceding form, *Nature* 246 (1973) 2226.