Multilevel capacities for the codon mutation channel

Dawit Nigatu and Werner Henkel
Jacobs University Bremen
Computer Science and Electrical Engineering
Email: d.nigatu@jacobs-university.de & werner.henkel@ieee.org

***Abstract*—We used the so-called empirical codon mutation (ECM) matrix to get some insights into the error protection mechanisms in the genetic code. First, using a dimension reduction technique, the codon substitution frequencies and the physicochemical properties of the encoded amino acids are compared. Then, the different levels of error protection in the genetic code are investigated by employing a 4-ary set partitioning scheme. The channel capacities of the set partitions and the level capacities were determined using the ECM channel model. One of the most interesting result which emanated from the capacity analysis is that the capacities of the first two partition levels were found to be identical. This implies that, in a multilevel coding sense, codes of equal code rates would have to be used to encode the two positions. The last codon position carries very little information and can be transmitted without coding.**

I. Introduction

The genetic code, which maps triplets of nucleotides (codons) to amino acids, is nearly universal [1], [2]. The mapping of 64 codons to 20 amino acids and a start and stop signals can be seen in Fig. 1. Multiple codons can encode for a single amino acid (synonymous codons) and due to this property, the code is described as degenerate. The properties, origin, and evolution of the genetic code have been extensively studied since its discovery in early 1960s [3], [4]. The code was shown to be optimal in terms of minimizing translation errors [5]–[7]. The optimality is determined by comparing the canonical genetic code to randomly generated codes which are obtained by shuffling the amino acid assignments, keeping the codon block structure. Usually, a metric that weights errors according to the effect they have on the structure and function of the synthesized protein is used to assess the optimality. Only a very small fraction of the random codes (one in a million [7]) outperformed the natural genetic code. In addition, the code is also optimized in regard to point mutations and frameshift errors [8]. Here, we like to examine the error protection capabilities of the code.

The empirical codon mutation (ECM) matrix, proposed by Schneider et al. [9] in 2005, shows rates of substitution between codons. The matrix is estimated from 8.3 million aligned codons from five vertebrate genomes which capture about 300 million years of evolutionary history. Unlike the models of evolution at nucleotide and protein levels, this codon-based model keeps the constraint imposed by the genetic code (the codon structure) and additionally shows substitutions between synonymous codons which would otherwise be veiled in protein-based models. In a previous work [10], we regarded the ECM matrix as a discrete memoryless communication channel and computed the channel capacity and determined an exponent with which the genetic information is reliably transmitted. In addition, we have shown the striking similarities between the optimal capacity achieving codon distribution and the one observed in real genomic sequences. In this paper, we analyzed the ECM matrix further with the aim of some more understanding of the error protection mechanisms in the genetic code. The codon substitution frequencies are compared to the physicochemical properties of the amino acids. Furthermore, the possibilities of a set partitioning scheme using the genetic code mapping is explored.

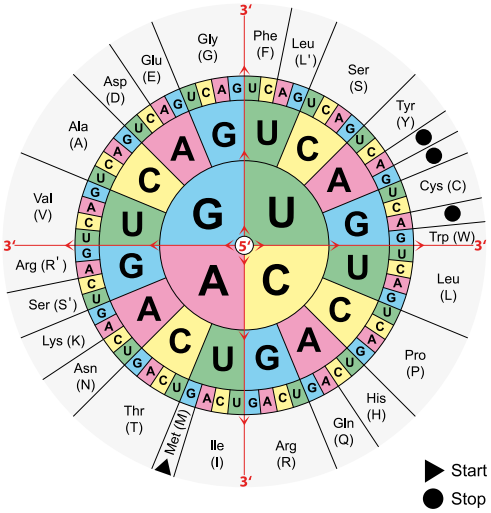

Fig. 1. The genetic code [11]. It is read from 5' to 3'. The primed amino acid symbols (R', S', and L') are similar to the unprimed ones and are labeled as such for illustration convenience.

The rest of this paper is organized as follows. First, in Section II, we present the relationship between mutation and chemical distances. Then, we proceed with performing capacity computations under a set partitioning of codon mappings in Section III. Finally, the main findings are summarized and concluding remarks are presented in Section IV.

II. Mutation vs. Chemical distances

When a substitution error occurs, the encoded amino acid can be the same (silent mutation), mostly, a substitution in the 3rd codon position results in a silent mutation, or it may produce a different amino acid (missense mutation). It can also lead to an early termination of the growing amino acid chain if the mutation is to a stop codon. The ECM matrix shows the

frequencies of substitutions between codons. In this paper, we would like to investigate the relationships between mutations and the difference or similarity in chemical properties of the encoded amino acids. The amino acid chemical distance matrix proposed by Grantham [12] is used as a measure of chemical similarity. Grantham's 20×20 amino acid chemical distance matrix was estimated from three chemical properties: composition, molecular volume, and polarity. Using the three properties as axes in Euclidean space, the distance ($D_{ij}^{(c)}$) between the ith and jth amino acid is computed. To perform the comparison, a similar mutation distance matrix has to be derived from the ECM matrix. To relate the mutation probabilities (P_{ij}) to distances between codons, we used the pairwise error probability (PEP) expression for a Gaussian i.i.d "channel" with a constant standard deviation (σ). The PEP is given by

$$P_{ij} = \frac{1}{2}\text{erfc}\frac{D_{ij}^{(m)}}{\sqrt{2}\sigma} \ , \tag{1}$$

where $D_{ij}^{(m)}$ is the Euclidean distance between ith and jth codons. The 61×61 mutation distance matrix is then obtained as

$$D_{ij}^{(m)} = \sqrt{2}\sigma\text{erfc}^{-1}(2P_{ij}) \ . \tag{2}$$

Due to the high dimensionality of the matrices a visual comparision is very difficult. Hence, we have used a technique called classical multidimensional scaling (CMDS) [13] to reduce dimensionality of the distance matrices and obtain a two-dimensional (2-D) view.

The two-dimensional (2-D) plots of the mutation and chemical distance matrices are shown in figures 2 and 3, respectively.

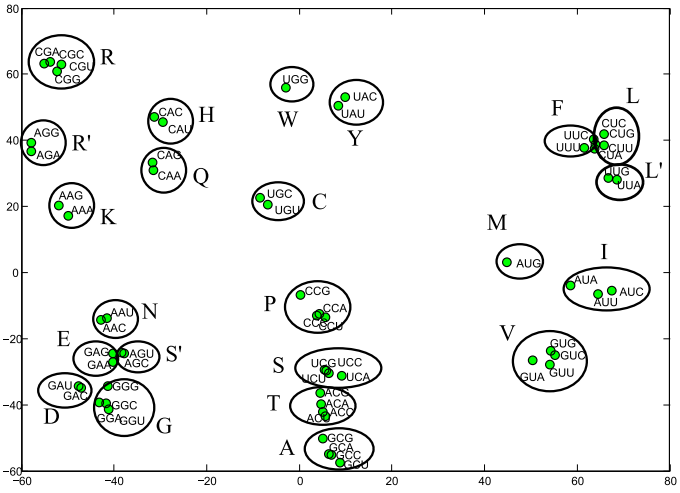

Fig. 2. 2-D plot of the mutation distance matrix.

From the 2-D plots, we observe that mutations between synonymous codons, encoding the same amino acid, are very high. They are bundled together. Likewise, chemically similar amino acids are close to each other. This reaffirms that highly probable mutations between codons lead to a chemically similar amino acid. In addition, the amino acid clusters are

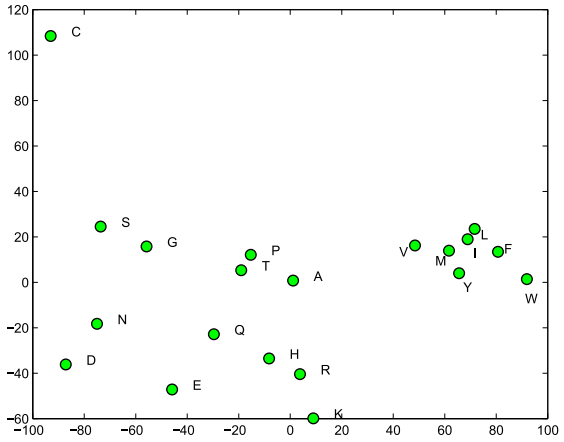

Fig. 3. 2-D plot of the chemical distance matrix.

mostly consistent with the so-called Taylor classification [14] shown in Fig. 4, which classifies amino acids based on their physicochemical properties. To mention an example, one can look at the hydrophobic amino acids F, L, M, I, and V in Fig. 2. Since most codon substitutions will not induce a significant change in the chemical properties, the produced protein will only be slightly affected and to some extent can still perform the intended function. This property further confirms the robustness of the genetic code to substitution errors [5]. However, there are some interesting inconsistencies where mutations are highly probable between amino acids having a relatively higher chemical distance. The observed inconsistencies are listed below.

Large chemical distance but small mutation distance:

- C with "all others"
- G with E
- S with {P,T,A}
- {D,N} with E
- {D,N} with G
- {Q,H} with {W,Y}
- K with N

Small chemical but large mutation distances are also observed ({W,Y} with {F,L,M,I,V} and {P,T,A} with {Q,H,R}) but these are not detrimental. The inconsistencies can be partly

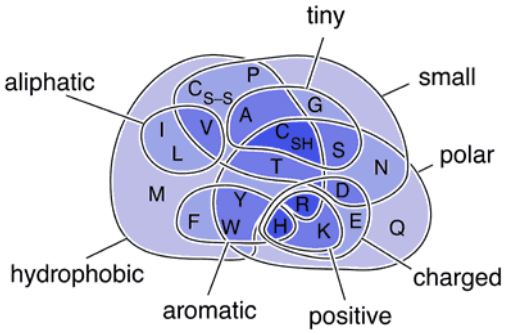

Fig. 4. Taylor classification of amino acids [15].

explained by a detailed study of the chemical properties. For

instance, cysteine (C), which allows for disulfide bridges, is critical for the stabilization of the protein structure [16] and, hence, is an outlier in the chemical distance matrix. In terms of mutation distance, however, cysteine is closer to multiple amino acids. These cases where small mutation distance but higher chemical distance require another level of error protection.

The existence of a genomic error correcting codes has been investigated by several researchers [17]–[20]. Gérard Battail [18], after showing that the capacity of the genetic channel vanishes exponentially with time, argued that to ensure the reliable transmission of genetic information, be it in the numerous replication cycles or over the geological time scale, can only be ensured with an intrinsic error correction mechanism. He further hypothesized that the codes used by nature are nested, i.e., concatenated codes. The nested coding structure indicates that nucleotides are unequally protected against errors, depending on their location. Hence, in the following, we will present a capacity analysis of the universal genetic code mapping employing a set partitioning scheme using the genetic channel described by the ECM matrix.

III. Set Partitioning of the genetic code

Coded modulation, which involves the joint optimization of coding and modulation, was independently presented by Ungerboeck [21], [22] and Imai and Hirakawa [23]. Ungerboeck introduced a mapping by set partitioning in which the modulation alphabet is successively partitioned into subsets in such a way that the minimum intra-subset Euclidean distances are maximized. Imai's approach of multilevel coding is to protect bits at different partition levels with separate codes. At the receiver end, typically multi stage decoding is performed [24]. Here, we used a 4-ary set partitioning (by virtue of the DNA/RNA alphabet size) and analyzed the capacities of the set partitions with the aim of understanding the universal genetic code mapping and the code constraints of such a system.

Since the closest points in the codon "constellation" (Fig. 2) are either synonymous or codons of a chemically similar amino acids, unlike Ungerboeck's original scheme, a block partitioning is preferred. The subsets are partitioned based on the three codon positions. Among the three positions, the 2nd is the most informative [25]. With the exception of Tryptophan (W), the most hydrophobic amino acids have U in the middle, while the most hydrophilic amino acids have A in the middle. Hence, in terms of keeping the similar amino acids together in the set partitions, starting from the 2nd base seems logical and indeed ensures the desired block partitioning. The 3rd position provides the least information in specifying the amino acids. Half of the time, the 3rd base contains no information. Therefore, it is used to partition at the last step. Figure 5 shows the partitions.

The codon-based ECM "channel" is specified by the 61×61 conditional (transition) probability $\mathbf{P}(y|x)$, where $x, y \in \mathcal{A}$. The set $\mathcal{A}$ contains all the 61 protein coding codons. The channel capacity is defined as the maximum mutual information between the transmitted codon X, $x \in \mathcal{A}$, and the received codon Y, $y \in \mathcal{A}$, over all possible input codon distributions, i.e.,

$$C = \sup_{P_X(x)} I(Y;X) \, . \tag{3}$$

The optimization problem was solved using the Arimoto-Blahut algorithm [26], [27]. To show the capacities at different "signal-to-noise ratio (SNR)" values, an exponent is introduced to the ECM matrix. In the ECM matrix, every codon has, on average, a 65 % mutation. Decreasing the exponent reduces the mutation rate and, hence, provides a better channel.

The mutual information $I(Y;X)$ is equivalent to $I(Y;X_1,X_2,X_3)$, where $x_1, x_2, x_3 \in \{C,U,A,G\}$. Applying the chain rule of mutual information [28], $I(Y;X)$ can be written as

$$\begin{aligned} I(Y;X) &= I(Y;X_1,X_2,X_3) \\ &= I(Y;X_2) + I(Y;X_1|X_2) + I(Y;X_3|X_1,X_2) \, . \end{aligned} \tag{4}$$

At the 1st partition level, using the 2nd base, the four subsets are $\mathcal{A}0 = \mathcal{A}(x_2 = C)$, $\mathcal{A}1 = \mathcal{A}(x_2 = U)$, $\mathcal{A}2 = \mathcal{A}(x_2 = A)$, and $\mathcal{A}3 = \mathcal{A}(x_2 = G)$. $\mathcal{A}0$ includes the codons encoding the amino acids $\{P,S,T,A\}$, all of them are tiny and small amino acids. $\mathcal{A}1 = \{L,F,M,I,V\}$ is a subset of codons encoding the hydrophobic amino acids, $\mathcal{A}2 = \{Q,H,Y,K,N,E,D\}$ are all polar, and $\mathcal{A}3 = \{R,W,C,S',G\}$ contains a diverse group of amino acids. For each subset, the corresponding sub-channel capacity can be calculated. To do so, we extracted the sub-matrix and re-normalized it so that each row sums to one. However, the input distribution is optimized only for the overall channel capacity and then the distribution of the subsets is selected and re-normalized as well. The capacities of the four sub-partitions along with the overall capacity are shown in Fig. 6. The weighted average is also shown. Except for the sub-group $\mathcal{A}0$, the capacities of the others decrease slowly with increasing exponent. Even for larger exponents (small "SNR"), the capacity is significant. However, the capacity of $\mathcal{A}0$ decreases very quickly (to zero) as the channel gets worst. This shows that the sub-group $\{P,S,T,A\}$ should also contain a relatively smaller information compared to the others.

At the 2nd partition level, using the 1st nucleotide, $\mathcal{A}0$ is subdivided into four distinct amino acids, which means that the amino acids are completely specified by the 1st two bases. The other sub-partitions do not provide a pure subset and, hence, the two codon positions are not sufficient to uniquely specify the amino acids (a third split is needed). The capacity curves of the subsets are depicted in Fig. 7. What is interesting to see from these plots is if the sub-channel is between synonymous codons, the capacities quickly vanish with increasing exponent (evolutionary time). This is because the inter-distances of the synonymous codons are too small (even zero, in some cases, see Fig. 2). However, when the sub-partitions are composed of a mixture of amino acids, the inter-distances are significant and provide non-zero capacities. Another interesting observation is, although 'W' and 'C', two chemically very different amino acids, are grouped together and the capacity of the {W, C} subset only slightly decreases with increasing exponent. This

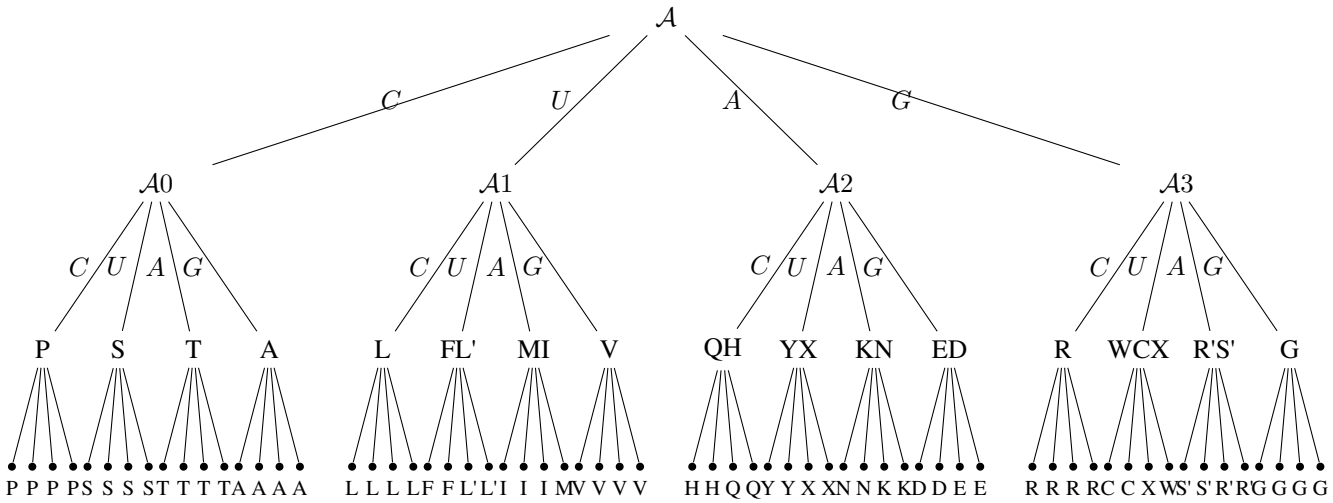

Fig. 5. Set-partitioning of the genetic code. $\mathcal{A}$ is a set containing all the 64 codons. Last two levels are labeled using the amino acid symbols. The symbol X represents the stop codons.

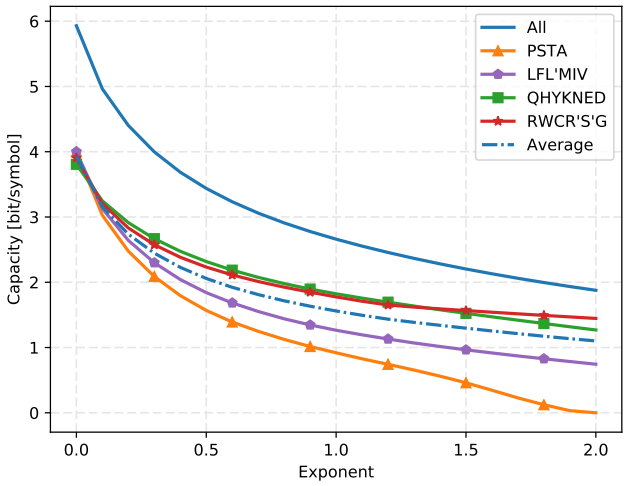

Fig. 6. Channel capacities at the first partition level.

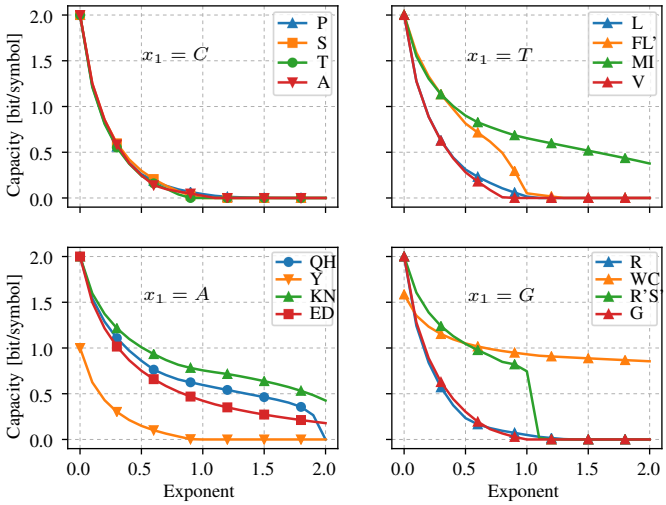

Fig. 7. Channel capacities at the second partition level.

indicates that the code or mapping would require to offer less protection/ higher rate.

The level capacity C^1 of the 1st partition level is calculated by, as a consequence of the chain rule, the difference between the mutual information before and after partitioning.

$$C^1 = I(Y;X_2) = I(Y;X_1,X_2,X_3) - I(Y;X_1,X_3|X_2)\ . \tag{5}$$

$I(Y;X_1,X_3|X_2)$ is obtained by averaging over all values of $x_2 \in \{C,T,A,G\}$. Similarly, the capacity of the 2nd partition level, C^2, can be calculated as,

$$C^2 = I(Y;X_1|X_2) = I(Y;X_1,X_3|X_2) - I(Y;X_3|X_1,X_2)\ , \tag{6}$$

where $I(Y;X_3|X_1,X_2)$ is computed by averaging over all possible combinations of x_1 and x_2.

$$I(Y;X_3|X_1,X_2) = \mathbb{E}_{x_1,x_2}\{I(Y;X_3|x_1,x_2)\}\ . \tag{7}$$

Note that, $I(Y;X_3|X_1,X_2)$ is the 3rd level capacity C^3.

The level capacities are shown together with the overall capacity and the average capacities of the set partitions (1st and 2nd) in Fig. 8. The level capacities, C^1 and C^2, provide the same capacity for the whole range. Hence, in this set partitioning scheme, the first two bases need to be protected by "codes" of a similar code rate. Around the exponent 1.25, the channel only permits an error-free transmission of 1 bit of information. This by itself is an interesting aspect which needs an explanation. The question is, how a symbol size of 4 can be transmitted using a capacity of 1 bit of information? One way is to use a clever mapping which merges two symbols as one. Although all four symbols are used, the actual information content can be 1 bit. This could, e.g., be achieved by classifying the fours symbols as purines or pyrimidines. However, the mappings should be in such a way that at least

the function of the produced protein is preserved. The other alternative is to use an error correcting code having a code rate specified by C^1 and C^2. When the capacity is too small, e.g., between the synonymous codons, error correction is not applied at all, but the mapping is "adjusted" to the too low capacity, i.e., extremely low capacities are not used at all. Otherwise, there must be an error correction or protection mechanism. What our result suggests is, if there is an error correcting code in the genome, the codes used to protect the two nucleotide positions, 1 and 2, should be the same.

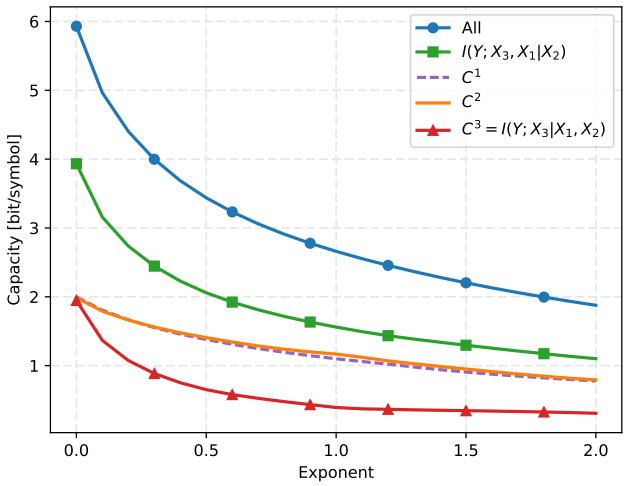

Fig. 8. Level capacities.

IV. Conclusion

A comparison between codon mutation distances and the respective amino acid chemical distances substantiates, with some exceptions, that highly probable mutations are either between synonymous codons or to chemically similar amino acids. The observed exceptions showing highly probable mutations between chemically very different amino acids suggest that there is a hidden protection mechanism. The different layers of error protection existing in the genetic code were investigated by applying a 4-ary set partitioning scheme. The ECM matrix was used as a channel model and an exponent was introduced to present the performances at different evolutionary times. Block partitioning of the codons was obtained by partitioning the codon set starting from the second nucleotide position and proceeding with the first and the third. An interesting aspect which emerged from the capacity analysis is that if a code (multilevel) would be assumed protecting the codon positions separately, the first two bases need to be encoded by codes having the same code rate. The capacities of the third partition level is often so small that it is not used at all, explaining the synonymous codons.

Acknowledgment

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – HE 3654/23-1.

References

[1] R. T. Hinegardner and J. Engelberg, "Rationale for a universal genetic code," *Science*, vol. 142, no. 3595, pp. 1083–1085, 1963.
[2] C. R. Woese, R. T. Hinegardner, and J. Engelberg, "Universality in the genetic code," *Science*, vol. 144, no. 3621, pp. 1030–1031, 1964.
[3] F. Crick, L. Barnett, S. Brenner, and R. J. Watts-Tobin, "General nature of the genetic code for proteins," 1961.
[4] M. W. Nirenberg, O. Jones, P. Leder, B. Clark, W. Sly, and S. Pestka, "On the coding of genetic information," in *Cold Spring Harbor Symposia on Quantitative Biology*, vol. 28. Cold Spring Harbor Laboratory Press, 1963, pp. 549–557.
[5] E. V. Koonin and A. S. Novozhilov, "Origin and evolution of the genetic code : the universal enigma," *IUBMB Life*, vol. 61, no. 2, pp. 99–111, 2012.
[6] D. Gilis, S. Massar, N. J. Cerf, and M. Rooman, "Optimality of the genetic code with respect to protein stability and amino-acid frequencies." *Genome biology*, vol. 2, no. 11, 2001.
[7] S. J. Freeland and L. D. Hurst, "The genetic code is one in a million," *Journal of Molecular Evolution*, vol. 47, no. 3, pp. 238–248, 1998.
[8] K. Mir and S. Schober, "Investigation of genetic code optimality for overlapping protein coding sequences," in *2014 8th Int. Symp. on Turbo Codes and Iterative Inf. Proc. (ISTC)*. IEEE, Aug 2014, pp. 152–156.
[9] A. Schneider, G. M. Cannarozzi, and G. H. Gonnet, "Empirical codon substitution matrix," *BMC bioinformatics*, vol. 6, no. 1, p. 134, 2005.
[10] D. Nigatu, A. Mahmood, and W. Henkel, "The empirical codon mutation matrix as a communication channel," *BMC bioinformatics*, vol. 15, no. 1, p. 80, 2014.
[11] W. Commons. (2009) Codons sun. [Online]. Available: https://commons.wikimedia.org/wiki/File:Aminoacids_table.svg
[12] R. Grantham, "Amino acid difference formula to help explain protein evolution," *Science*, vol. 185, no. 4154, pp. 862–864, 1974.
[13] I. Borg and P. Groenen, "Modern multidimensional scaling: theory and applications," *Journal of Educational Measurement*, vol. 40, no. 3, pp. 277–280, 2003.
[14] W. R. Taylor, "The classification of amino acid conservation," *Journal of theoretical Biology*, vol. 119, no. 2, pp. 205–218, 1986.
[15] W. Commons. (2007) Venn diagram of amino acids. [Online]. Available: https://commons.wikimedia.org/wiki/File:Amino Acids Venn Diagram.png
[16] M. J. Betts and R. B. Russell, "Amino acid properties and consequences of substitutions," *Bioinformatics for geneticists*, vol. 317, p. 289, 2003.
[17] G. Battail, "Genomic error-correcting codes in the living world," *Biosemiotics*, vol. 1, no. 2, pp. 221–238, 2008.
[18] ——, "An outline of informational genetics," *Synthesis Lectures on Biomedical Engineering*, vol. 3, no. 1, pp. 1–205, 2008.
[19] L. S. Liebovitch, Y. Tao, A. T. Todorov, and L. Levine, "Is there an error correcting code in the base sequence in DNA?" *Biophysical Journal*, vol. 71, no. 3, pp. 1539–1544, 1996.
[20] E. E. May, M. A. Vouk, D. L. Bitzer, and D. I. Rosnick, "An error-correcting code framework for genetic sequence analysis," *Journal of the Franklin Institute*, vol. 341, no. 1-2, pp. 89–109, 2004.
[21] G. Ungerboeck, "Channel coding with multilevel/phase signals," *IEEE Transactions on Information Theory*, no. 1, pp. 55–67, Jan.
[22] ——, "Trellis-coded modulation with redundant signal sets Part I: Introduction," *IEEE Communications Magazine*, no. 2, pp. 5–11, feb.
[23] H. Imai and S. Hirakawa, "A new multilevel coding method using error-correcting codes," *IEEE Transactions on Information Theory*, vol. 23, no. 3, pp. 371–377, 1977.
[24] U. Wachsmann, R. F. Fischer, and J. B. Huber, "Multilevel codes: Theoretical concepts and practical design rules," *IEEE Transactions on Information Theory*, vol. 45, no. 5, pp. 1361–1391, 1999.
[25] F. Taylor and D. Coates, "The code within the codons," *Biosystems*, vol. 22, no. 3, pp. 177–187, 1989.
[26] S. Arimoto, "An algorithm for computing the capacity of arbitrary discrete memoryless channels," *Information Theory, IEEE Transactions on*, vol. 18, no. 1, pp. 14–20, Jan 1972.
[27] R. Blahut, "Computation of channel capacity and rate-distortion functions," *Information Theory, IEEE Transactions on*, vol. 18, no. 4, pp. 460–473, Jul 1972.
[28] T. M. Cover and J. A. Thomas, "Elements of information theory 2nd edition," 2006.