Proceedings of the Indiana Academy of Science
(1987) Volume 97 p. 133-136.

An Optimized Genetic Code

T. Alvager, G. Graham, D. Hutchison and J. Westgard
Departments of Life Sciences, Mathematics and Computer Science and Physics
Indiana State University, Terre Haute, Indiana 47809

Introduction

In living organisms 20 amino acids are encoded by 64 codons. These numbers are often regarded as pure coincidences determined at random in some remote epoch when the genetic machinery of the cell was finalized [see e.g. refs 1 and 2]. However, it is not excluded that a deterministic mechanism is working and that, for instance, the number of coded amino acids is determined by the genetic code itself or its environment and therefore a systems theoretical approach may be appropriate.

The first investigator to consider the latter possibility in some detail seems to have been Garnow [3]. He derived the magic number 20 from a special model of the genetic apparatus. This particular model was later found to be less satisfactory and Garnow's derivation is therefore most likely not acceptable today.

A recent calculation of the correct number of the coded amino acids has been given by Soto and Toka [4]. This theory is based on a principle that defines the optimal code as one that minimizes for a given amount of information encoded, the product of the number of physical devices used by the average complexity of each device. The theory obtains the number 20, but does not give an explanation of the degeneracy and partitioning of the codons, which play such an important part in the genetic code problem.

In this work we present a derivation of the numbers involved in the genetic code problem that is based on the degeneracy of the codons and an assumption that the cell functions under the pressure of delivering an effective product. It is found that not only can the number of amino acids be derived in a simple manner but also the actual partitioning of degeneracy of the codons can be clarified.

Theory

The basic assumption of the theory is that the genetic code machinery of a cell is functioning under the constraint of optimizing its operations. Empirical genetic code data must be used as a guide to find the proper function to optimize.

Table 1 gives the genetic code in the so called universal or standard form (SGC) with the usual abbreviations for basis and amino acids [see e.g. ref 5]. It should be noted that the codons are divided into partitions according to the degeneracy of the codons

Table 1. The standard genetic code (see e.g. ref 5)

Leu UUA	Ser UCU	Gly GGU	Pro CCU	Tyr UAU	Ter UAA
UUG	UCC	GGC	CCC	UAC	UAG
CUU	UCA	GGA	CCA	Glu GAA	
CUC	UCG	GGG	CCG	GAG	Ile AUU
CUA	AGU	Ala GCU		Cys UGU	AUC
CUG	AGC	GCC	Asn AAu	UGC	AUA
Arg CGU		GCA	AAC	His CAU	
CGC	Thr ACU	GCG	Lys AAA	CAC	Met AUG
CGA	ACC	Val GUU	AAG	Aln CAA	Trp UGG
CGG	ACA	GUC	Asp GAU	CAG	Ter UGA
AGA	ACG	GUA	GAC	Phe UUU	
AGG		GUG		UUC	

with each partitioning corresponding to an amino acid or a terminator. Thus, for instance, the partition block for the amino acid Ser consists of six codons. This will be referred to as a 6-degeneracy. The various degeneracies are shown in Table 1.

The theory will be discussed in terms of a set of variables, $x_1, x_2, \ldots x_i, \ldots x_n$, where x_i gives the number of partitions of a particular degeneracy i. If there are a total of N possible codons we have the general constraint

$$\sum_{i=1}^{n} i\, x_i = N \qquad (1)$$

with n different partitionings.

In the case of the SGC, the number N is 64 since there are 64 codons generated from the alphabet of the four letters U, C, A and G arranged in groups of three. Here as usual U,C,A and G represent the four bases uracil, cytosine, adenine and guanine.

The standard genetic code as displayed in Table 1 shows clearly that partitions with an even number of degeneracies are favored in nature. In fact only three partitions occur for i = 1 and one partition with i = 3. This rarity of partitions with odd numbers is even more pronounced in some of the genetic codes that have been discovered recently to occur in mitochondria [see e.g. ref 6].

The predominance of partitions with an even number of degeneracies is most likely due to the physical character of the molecules involved. Thus if the purpose is to consider the genetic code from a systems point of view it seems reasonable to accept as a starting point the basic empirical fact of predominance of partitionings of even degeneracy. This argument is also supported by symmetry considerations using group theoretical calculations [7].

An inspection of the date in Table 1 shows that $2x_2(=20)$ is equal to $4x_4(=20)$ which in turn is approximately equal to $6x_6(=18)$. This suggests that the function to optimize should be a simple product function.

The simplest function that is reasonable to apply is therefore a multilinear function, C, given by

$$C(x_1 \ldots x_n) = k \prod_i x_i \qquad (2)$$

with k being a constant. The problem is then to optimize the function $C(x_1 \ldots x_n)$ with the contraint of eq (1). By using the Lagrange multiplier method with constant β the problem can be formulated as an optimization of a function F given by

$$F = k \prod x_i - \beta \sum i x_i \qquad (3)$$

From the conditions $\partial F/\partial x_i = 0$ (i = 1,...n) we therefore obtain the following system of equations:

$$\begin{aligned} &k \prod_{i \neq m} x_i = \beta m \qquad (m = 1 \ldots n) \\ &\sum i\, x_i = N \end{aligned} \qquad (4)$$

to determine the n parameters x_i

Results and Discussion

Before solving the system (4) it is convenient to introduce the approximation stated in the previous section consisting of neglecting terms with odd degeneracy. We also use the empirical fact that there are no known eight-fold degeneracies. Thus specializing to the case N = 64, and i = 2, 4 and 6 the system of equations (4) reads:

$$\begin{aligned} k\, x_4\, x_6 &= 2\beta \\ k\, x_2\, x_6 &= 4\beta \\ k\, x_2\, x_4 &= 6\beta \\ x_2 + 2x_4 + 3x_6 &= 32 \end{aligned} \qquad (5)$$

This system has the following set of solutions:

$$\begin{aligned} x_2 &= 10.68 \\ x_4 &= 5.34 \\ x_6 &= 3.56 \end{aligned} \qquad (6)$$

This corresponds to 11 partitions with 2-degeneracy, 5 partitions with 4-degeneracy and 4 partitions with 6-degeneracy. The total number of partitions is 20.

An examination of the SGC as displayed in Table 1, shows that the code has 10 partitions with 2-degreneracy, 5 partitions with 4-degeneracy and 3 partitions with 6-degenercy. There is also 1 partition with 3-degeneracy and 3 partitions with 1-degeneracy.

The theoretical model is therefore in good agreement with the SGC partitions. This is especially true if the originally calculated numbers i.e. $x_2 = 10.68$, $x_4 = 5.34$ and $x_6 = 3.56$ are taken into account in the comparison and not just the rounded off, integer numbers.

One can also compare the prediction of the model with the genetic code as expressed in mitochondria of some species having a genetic code differing from SGC. Thus, for instance, in Protozoa [6] the genetic code has changed somewhat from SGC in that the codon UGA codes for the amino acid trp and is not a terminator. This changes the pattern of partitions into 11, 5 and 3 partitions for 2-,4- and 6-degeneracy respectively, which is in excellent agreement with the model values.

Table 2 gives a summary of experimental values and the predictions of the model.

Table 2. Number of degeneracies in SGC, some mitochondrial genetic codes and the corresponding model values (numbers in parantheses are the actual calculated values). Mitochondrial data are from Breitenberger [6]

Object	1-degen.	2-degen.	3-degen.	4-degen.	6-degen.
SGC	3	10	1	5	3
Mammals	0	14	0	6	2
S. cerevisiae	1	11	1	5	3
N. crassa	1	11	1	5	3
Protozoa	1	11	1	5	3
Model	—	11(10.68)	—	5(5.34)	4(3.56)

The general trend indicated by the data is in good agreement with the model. Other expressions than the one given by relation (2) have been tried. However, the tabulated relation seems to be the optimum one.

Conclusion

Most biological structures are determined by physical as well as system contraints. In the present work it is assumed that the physical constraint on the genetic code determines the basic 2-degeneracy. However, this degeneracy is modified by system constraints in such a way that some higher degeneracy is favored. The simple function that has been used seems to explain the gross structure of the existing partitions. The exact biological interpretation of the function is left open at this time. This procedure is therefore similar in spirit to the group theoretical arguments given by Findley and Mc Glynn [7] in that a mathematical structure is assumed without first introducing a particular biological func-

tion. Future work on the model must obviously address the interpretation question as well as the question concerning the actual assignment of codons to a particular amino acid. Such a task will most likely require additional physical as well as evolutionary considerations.

Literature Cited

1. Eigen, M., W. Gardiner, P. Schuster and R. Winkler-Oswatitsch, 1981. The origin of genetic information. Sci. Am. 244: 78-94.
2. Jukes, T.H., 1983. In M. Nei and R. Koehne (eds) *Evolution of the amino acid code in Evolution of genes and proteins*, Sinauer Ass., 191-207.
3. Garnow, G. 1954. Possible relations between Deoxyribonucleic acid and protein structure. Nature, 173: 318.
4. Soto, A. and J. Toka, 1985. A hardware interpretation of the evolution of the genetic code. BioSystems, 18: 209-215.
5. Lehninger, A., 1974. *Biochemistry*, Worth Publ.
6. Breitenberger, C. and U. Raj Bhandary, 1985. Some highlights of mitochondrial research based on analyses of Neurospora crassa mitochondrial DNA, TIBS December 1985, 478-482.
7. Findley, G.L. and S.P. McGlynn, 1979. A generalized genetic code. International Journal of Quantum Chem., 6: 313-327.