Standard Genetic Code Degeneracies from Maximum Information Calculations

T. Alvager,*,†,‡ G. Graham,§ D. Hutchison,§ and J. Westgard†

Department of Physics, Department of Mathematics and Computer Science, and Interdisciplinary Center for Cell Products and Technologies, Indiana State University, Terre Haute, Indiana 47809

Received October 21, 1993®

In the known biological codes, which include the standard genetic code and various mitochondrial codes, 20 amino acids along with terminator value(s) are encoded by 64 codons resulting in certain degeneracy patterns for these codes. The present work offers a model based on a generalized information function designed to predict these degeneracy patterns. Previous work using this approach has had some success with the mitochondrial codes, showing the importance of a random component in the determination of the degeneracies. This paper extends the method to the degeneracy pattern of the standard genetic code.

INTRODUCTION

The experimental facts about the genetic code are well-known.[1-3] Yet, in spite of considerable theoretical work,[4-9] certain aspects of its structure are still not completely understood. In particular the degeneracy patterns associated with the various codes have no quantitative explanation.[10] All codes use 64 codons and code only for 20 amino acids plus some terminators. Consequently most amino acids correspond to more than one codon. It is not obvious, however, why the degeneracy distributions are the ones found experimentally.

Two of the experimentally known degeneracy patterns for genetic codes, including the standard genetic code, are listed in Table 1. It should be noted that the codes have two terminator groups (UAA and UAG) and (UGA). The total number of code groups is therefore 22. A variety of other codes are known.[3] However, for our purpose the two listed codes are the important ones.

In this paper we address this circumstance with the degeneracy distributions and suggest a phenomenological two-parameter information function to calculate the degeneracy distribution for genetic codes, in particular, for SGC. We obtain the known degeneracy patterns for certain parameter values. This is in accordance with previous work in which this approach, but with only one parameter, has had some success with the mitochondrial codes.[12] We emphasize that only the degeneracy problem is analyzed in this paper. The question of how the codons are assigned to particular amino acids is not discussed.

NOTATION

It is convenient to introduce the following notation. Let all codons that code for the same amino acid be in a single block. A block with k codons is said to be k-degenerate. Denote the number of k-degenerate blocks by x_k. Then, the code can be labeled

$$x = (x_1, x_2, ..., x_k, ..., x_m) \quad (1)$$

where m is the largest k for which x_k is different from 0. Since there are 64 codons, we have the general constraint

$$\sum_k kx_k = 64 \quad (2)$$

Table 1. Number of Degeneracies in the Standard Genetic Code (SGC) and for Comparison the Mitochondrial Genetic Code in Mammals[a]

object	1-de	2-de	3-de	4-de	5-de	6-de	7-de	8-de	A
SGC	3	10	1	5	0	3	0	0	22
mammals[2]	0	14	0	6	0	2	0	0	22

[a] 1-de stands for 1-degeneracy, etc. The sum of all degeneracy numbers, including terminator values, are shown in the last column, labeled A; i.e., $A = \sum_k x_k$.

where the summation is taken from $k = 1$ to $k = m$. In this notation the experimental values for the standard genetic code can be represented by the x-distribution

$$X_{SGC} = (3,10,1,5,0,3) \quad (3)$$

In this distribution two blocks (UAA, UAG and UGA) of terminators are included with degeneracies 2 and 1, respectively.

A quantitative theory of the degeneracies of the genetic code should be able to predict the numbers specified in (3). It has been shown elsewhere[12] that a useful tool for describing certain features of the code is given by extreme values of a generalized information function of the form[13,14]

$$I_g = R + G \quad (4)$$

with

$$R = -\sum_k (kx_k/64)\{\ln(kx_k/64)$$

$$G = -\sum_k (kx_k/64)g(k)$$

where $g(k)$ is a function to be specified that describes a nonrandom effect. For $g(k) = 0$, expression 4 becomes the regular Shannon information function,[15,16] which is a measure of uncertainty about which m events occur from a partition of the sure event with known probabilities $p_k = kx_k/64$.

NUMERICAL COMPUTATIONS OF DEGENERACIES

A simple choice for the function $g(k)$ is given by

$$g(k) = -[\ln\{(\prod_n |k - 2n + 1|)^a + (\prod_n |k - 2n + 2|)^b\}] \quad (5)$$

where a and b are parameters to be determined and the

† Department of Physics.
‡ Interdisciplinary Center for Cell Products and Technologies.
§ Department of Mathematics and Computer Science.
® Abstract published in *Advance ACS Abstracts*, June 1, 1994.

0095-2338/94/1634-0820$04.50/0

Table 2. Maximum Value of the Function I_g, the Corresponding Distribution (x_1, x_2, ..., x_6), and $\sum_k x_k$

		a		
b		-0.5	0.0	0.5
-3.0	maximum I_g	0.39	1.09	1.99
	distribution	(1,12,1,6,0,2)	(0,10,0,5,0,4)	(0,8,0,3,0,6)
	$\sum_k x_k$	22	19	17
-2.0	maximum I_g	0.48	1.16	2.00
	distribution	(4,11,2,5,0,2)	(3,10,1,5,0,3)	(1,7,1,4,0,5)
	$\sum_k x_k$	24	22	18
-1.0	maximum I_g	0.73	1.30	2.07
	distribution	(10,9,3,4,1,1)	(6,9,2,4,0,3)	(3,8,1,3,0,5)
	$\sum_k x_k$	28	24	20

Table 3. Maximum Value of the Function I_g, the Corresponding Distribution (x_1, x_2, ..., x_8), and $\sum_k x_k$

		a		
b		-0.5	0.0	0.5
-3.0	maximum I_g	-0.06	1.38	3.04
	distribution	(0,9,0,5,0,3,0,1)	(0,7,0,4,0,3,0,2)	(0,6,0,2,0,2,0,4)
	$\sum_k x_k$	18	16	14
-2.0	maximum I_g	-0.06	1.38	3.04
	distribution	(0,9,0,5,0,3,0,1)	(0,7,0,4,0,3,0,2)	(0,6,0,2,0,2,0,4)
	$\sum_k x_k$	18	16	14
-1.0	maximum I_g	0.16	1.42	3.04
	distribution	(3,7,2,4,1,2,0,1)	(1,8,1,4,0,2,0,2)	(0,6,0,2,0,2,0,4)
	$\sum_k x_k$	20	18	14

Table 4. Summary of Computations for Genetic Codes Listed in Table 1[a]

object	A	a	b	I_g	R	|G|
SGC	22	0.0	-2.0	1.2	1.0	0.2
mammals[2]	22	-0.6	-4.0	1.0	0.2	0.8

[a] Parameters a and b are calculated according to the scheme described in the text and Tables 2 and 3. $I_g = R + G$ is from eq 4.

products are taken from $n = 1$ to $n = m/2$. Note that if k is odd, the first product in (5) is 0 and if k is even, the second is 0. This choice takes into account the experimental fact that the odd degeneracies seem to form a group different from the even ones. The two parameters used in the code expression have therefore an important and natural interpretation. They refer to the relative contributions of even and odd kinds of degeneracies.

Table 2 lists the results from a computer search through all possible distributions (x_1, x_2, ..., x_m) for $m = 6$ that maximizes the generalized information function I_g. Note that we take a and b to have the listed values unless the corresponding product is 0 and a or $b \leq 0$. In this case the exponent is taken to be positive so that there is a zero contribution to $g(k)$, rather than an infinite one. The search shown in the table corresponds to a narrow range of values for a and b found through a successive fine tuning of a larger set of a and b values originally suggested by an analytical method to calculate degeneracies.[12] The computations indicate that the SGC distribution is obtained for $a = 0 \pm 0.1$ and $b = -2.0 \pm 0.3$. In the specified range the distribution gives the correct number of amino acids and terminators; i.e., $\sum_k x_k = 22$.

Table 3 lists corresponding results for $m = 8$. An 8-degeneracy could in principle be formed from a combination of one 6-degeneracy and one 4-degeneracy giving one 8-degeneracy and one 2-degeneracy, thus keeping the number of coded amino acids constant. In this case the data in Table 2 show no distribution that results in $\sum_k x_k = 22$ and a positive value for the I_g function. This result may be taken as evidence for the prediction from our model that there are no 8-degeneracies in SGC since for nearby codes (in the sense of close values for a and b) with $\sum_k x_k = 22$ the extreme values come out negative. Assuming that higher degeneracies have evolved in time, it is reasonable to surmise that once a number $\sum_k x_k = 22$ had been established, an increase above 6 in m was ruled out.

It should be noted, however, that there are codes containing an 8-degeneracy, as for instance the mitochondrial code in Nematodes.[11] This case is clearly different from the SGC in the fact that $\sum_k x_k = 21$. Since the emphasis in this work is on the randomness in the SGC, other cases will be discussed elsewhere. A summary of obtained results for the SGC code and mitocondrial code listed in Table 1 are presented in Table 4.

DISCUSSION

The phenomenological theory presented in this work allows a calculation of the SGC distribution. According to the model there are two contributing parts to the explanation of the distribution: a nonrandom element expressed in the function $g(k)$ and a random component. The function $g(k)$ should be possible to obtain from molecular reaction considerations resulting in a derivation of the parameters a and b. The important finding, however, in this work is the quantitative estimate of the random contributing part to the degeneracy. In the SGC the degeneracy distribution is the dominant one, while in the mammalian mitochondrial code the nonrandom part dominates.

REFERENCES AND NOTES

(1) Watson, J. D.; Hopkins, N. H.; Roberts, J. W.; Steitz, J. A.; Weiner, A. M. *Molecular Biology of the Gene*, 4th ed.; Benjamin/Cummings: Menlo Park, NJ, 1987; Vol. 1, Chapter 15, p 431.

(2) Fox, D. T. Natural variation in the genetic code. *Annu. Rev. Genet.* **1987**, *21*, 67–92.

(3) Osawa, S.; Jukes, T.; Watanabe, K.; Muto, A. Recent Evidence for Evolution of the Genetic Code. *Microbiol. Rev.* **1992**, *56*, 229–264.

(4) Crick, F. H. C. Codon-anticodon pairing. *J. Mol. Biol.* **1966**, *19*, 548–555.

(5) Wong, J. T. The evolution of a universal genetic code. *Proc. Natl. Acad. Sci. U.S.A.* **1976**, *73*, 2336–2340.

(6) Findley, G. L.; McGlynn, S. P. A generalized genetic code. *Int. J. Quantum Chem.* **1979**, *6*, 313–327.

(7) Eigen, M.; Gardiner, W.; Schuster, P.; Winkler-Oswatitsch, R. The origin of the genetic information. *Sci. Am.* **1981**, *244*, 78–94.

(8) Antillon, A.; Ortega-Blake, I. A group theory analysis of the ambiguities in the genetic code. *J. Theor. Biol.* **1985**, *112*, 757–769.

(9) Soto, A.; Toha, J. A hardware interpretation of the evolution of the genetic code. *BioSystems* **1985**, *18*, 209–215.

(10) Alvager, T.; Graham, G.; Hilleke, R.; Hutchison, D.; Westgard, J. On the information content of the genetic code. *BioSystems* **1989**, *22*, 189–196.

(11) Okimoto, R.; Macfarlane, J. L.; Clary, D. O.; Wolstenholme, D. R. The Mitochondrial Genom of Two Nematodes, Caenorhabditis elegans and Ascaris suum. Genetics **1992**, *130*, 471–494.

(12) Alvager, T.; Graham, G.; Hilleke, R.; Hutchison, D.; Westgard, J. 1990, A generalized information function applied to the genetic code. *BioSystems* **1990**, *24*, 239–244.

(13) Aczel, J.; Forte, B. Generalized entropies and maximum entropy principle. In *Maximum entropy and Bayesian methods in applied statistics*; Justice, J., Ed.; Cambridge Press: Cambridge, U.K., 1986; pp 95–100.

(14) Bevensee, R. M. *Maximum Entropy Solutions to Scientific Problems*; Prentice Hall: New York, 1993.

(15) Shannon, C.; Weaver, W. *The Mathematical Theory of Communication*; University of Ilinois Press: Urbana, IL, 1949.

(16) Gatlin, L. L. *Information Theory and the Living System*; Colombia University: New York, 1972.