The e-Dimensionality of Genetic Information

The *e*-Dimensionality of Genetic Information

Subhash Kak
Oklahoma State University, Stillwater

Abstract.
This paper provides an explanation for why the assignment of codons to amino acids, which range from 1 to 6, is non-uniform. Since mathematical coding theory demands a near uniform assignment, the answer to this question is important to understand deeper aspects of the structure of the genetic code. Our analysis points to 20 different covering regions in an *e*-dimensional information space, which is equal to the number of amino acids. It is also shown that the assignment of the codons to the amino acids is fractal-like that is well modeled by the Zipf distribution. It is remarkable that the Zipf distribution that holds for the letter frequencies of words in a language also applies to the rank order of triplets the code for amino acids.

Introduction

The three-letter standard genetic code [1][2][3][4] with 64 codons -- that include 3 stop triples -- has 61 codons that specify 20 amino acids. The code has been studied from different perspectives including representation in abstract multi-dimensional space to determine how it may have evolved [5][6][7].

The genetic code, the sequence of nucleotides in deoxyribonucleic acid (DNA) and ribonucleic acid (RNA), maps into the amino acid sequence of proteins. This is a two-step process where the strings of cytosine (C), guanine (G), adenine (A), and thymine (T) in groups of three in the DNA leads to the synthesis of a messenger RNA (mRNA) molecule, in which thymine (T) is replaced by uracil (U), which directs the formation of the protein.

The AUG codon, in addition to coding for methionine, is found at the beginning of every mRNA and indicates the start of a protein. Methionine and tryptophan are the only two amino acids that are coded for by just a single codon (AUG and UGG, respectively). The other 18 amino acids are coded for by two to six codons. Because most of the 20 amino acids are coded for by more than one codon, the code is degenerate.

In this paper, we consider aspects of its mathematical structure that have escaped earlier scrutiny. Specifically, we ask that since Nature is optimal why is the mapping of the codons into amino acids not uniform, and a related question is why only 20 amino acids, out of hundreds, are used as building blocks of proteins [8]. A uniform mapping would assign nearly the same number of codons to each amino acid which should thus be 61/20, or about 3. In reality, the assignment is highly variable as shown in Table 1 (summarizing information of Table 2) and Figure 1.

Table 1. Number of codons to amino acid

codons to amino acid	1	2	3	4	5	6
number of this type	2	9	1	5	0	3

The assignment ranges from 1 to 6 and we know from mathematical coding theory [9] based on integer-dimensional spaces that design of such a code is highly inefficient in its capacity to correct errors. To examine why Nature would choose such an assignment, we consider the question of optimal representation of information [10].

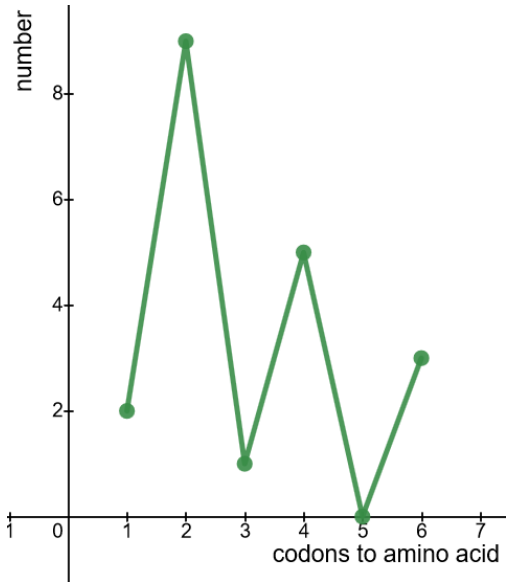

Figure 1. The number of codons assigned to amino acids varies from 1 to 6

The most efficient way to represent structural information is in *e*-dimensions [11][12][13]. We analyze this noninteger space from the perspective of the genetic code and provide a novel explanation for the redundancy [14] in the codon table.

Table 2. RNA codon table from the National Human Genome Research Institute

RNA codon table

1st position	2nd position U	2nd position C	2nd position A	2nd position G	3rd position
U	Phe	Ser	Tyr	Cys	U
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	stop	stop	A
	Leu	Ser	stop	Trp	G
C	Leu	Pro	His	Arg	U
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	U
	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys	Arg	A
	Met	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	U
	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A
	Val	Ala	Glu	Gly	G

Specifically, we find that the number of basic shapes needed to cover the *e*-dimensional information space is 20, which equals the number of amino acids associated with the genetic code. While this may be a coincidence, the possibility that information theoretic constraints are the explanation is compelling in view of other fractal-like characteristics of the code that will be described in the paper.

Preliminaries

Nature privileges optimality, therefore information structures in natural systems may be analyzed from the perspective of efficiency. If the most basic structures constitute the alphabet, their combination in different ways will yield more complex forms quite like the way the alphabet of a language leads to words, sentences and narratives.

The basic structures may be defined in terms of number or number in physical or an appropriate abstract information space. For example, the periodic table of chemical elements is a tabular display in which the elements are arranged by atomic number and electron configuration. The problem of finding the alphabet of forms that can cover the information space concerns us here.

The basic form can be defined by coordinates. The simplest physical geometry that when aggregated or divided can cover the 3-dimensional space is the cube. But we must go beyond the three-dimensional cube since the optimal geometry is *e*-

dimensional, and as *e* is closer to 3 than to 2, it follows that ternary logic is superior to binary logic [15][16] which is another mathematical fact of relevance to biology.

The proof the assertion that *e*-dimensionality is optimal is elementary. Each coordinate axis in the general abstract space may be viewed as a bin. Assume a total of *d* bins and label them as 1, 2, 3…*d*. The utilization of the system would be optimal if each of the bins carries the same information or the probability of the use of each is equal to 1/*d.* The information associated with each bin then equals $\ln d$.

This information increases as *d* increases, but this increase is obtained at the cost of the use of the larger and thus more expensive bin set. The information efficiency per bin is:

$$E(d) = \frac{\ln d}{d}$$

which is shown in Figure 2. Its maximum value is obtained by taking the derivative of *E(d)* and equating that to zero, which yields $d_{opt} = e = 2.71828....$

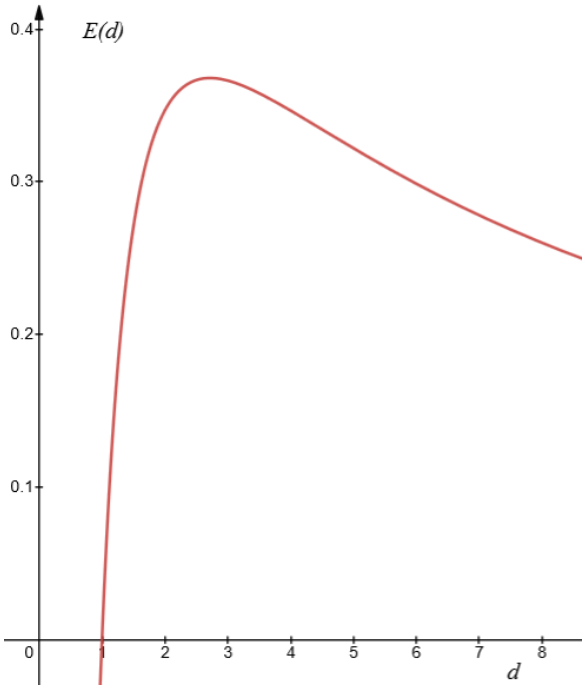

Figure 2. Efficiency of representation

In other words:

Theorem 1

The optimal number of bins associated with representation is e.

The bins may be viewed as the coordinate axes of the corresponding abstract space or be aggregated as logical classes. Another perspective is that this represents a fractal structure which means a space that is somehow has gaps in its structure. The noninteger nature of dimensionality may be seen in a complementary perspective as the source of an attractive force within the enclosing integer-dimensional space [17][18][19].

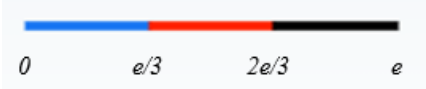

Figure 3. The division for one dimension

One can also view it from the perspective of the most efficient coding of data, in which case *e* represents the radix.

Two-dimensional case to cover the space

The basic shape is the square that may be divided progressively or aggregated to yield larger forms. The division into 4 sub-squares is to emphasize the need for smaller areas. Note that each axis requires three points, such as (0, ½,1), to locate the sub-squares.

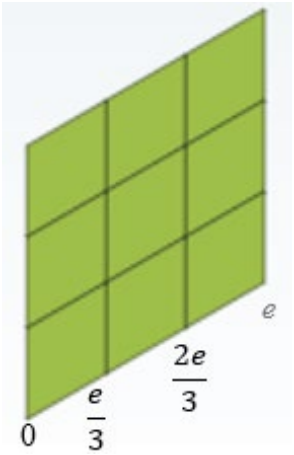

Figure 4. The 9 sub-squares on the 2-d surface

For the representation of a 2-d projection of a 3-d object, one will need an *e*-way division along each axis (Figure 4). On a two dimensional plane, the number of bins (sub-squares) will be:

$$e^2 \approx 7.389$$

This means that although there are 9 squares in the e-sided square with each of side *e*/3, the number from the perspective of unit-based observation is approximately 7.4, which may be rounded to 7.

Three-dimensional structure

Paralleling the previous cases, we consider a cube that is subdivided into smaller sub-cubes. The points at the intersection of the sub-cubes are 64 in total. If the information regions are the cubes, and the mapping is taken to be relational, then 27 sub-cubes will be defined.

Theorem 2

An e-dimensional cube with e-way division along each axis will have a total of $e^3 \approx 20$ *sub-cubes.*

Each of the sub-cubes has a volume of $\left(\frac{e}{3}\right)^3 = \frac{e^3}{27}$, and the 27 of them add up to the volume of e^3. However, from the perspective of a unit-based geometry, this volume will map to 20 sub-cubes that may be combined together to cover the larger space.

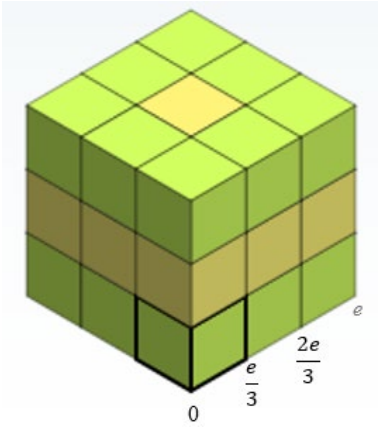

Figure 5. The 27 sub-cubes for the 3-dimensional space

Protein sequences consist of 20 commonly occurring amino acids; therefore, it can be said that the protein alphabet consists of 20 letters. Different amino acids have different chemistries (such as acidic versus basic, or polar and nonpolar) and different structural constraints. Variation in amino acid sequence is responsible for the enormous variation in protein structure and function. The 20 sub-cubes of the *e*-dimensional information space of the genetic code provide a different perspective on the genetic code alphabet.

Fractal characteristics of the genetic information space

Consider the rank order associated with the number of codons to amino acids which is shown in Table 3. The same pattern repeats to a different scale as we trace from left to right in Figure 1, which is the hallmark of a fractal or noninteger dimensional process [20][21][22][23][24].

The number of each type is also in a broad distribution with values ranging from 0 to 9 and is given in the rank order frequency as in Table 3:

Table 3. Number of codons to amino acid

rank order	1	2	3	4	5	6
frequency	9	5	3	2	1	0

This is approximately given by the Zipf's law:

$$f = 9r^{-b}, b = 1$$

In this small set, the divergence of the actual frequencies is quite small since the theoretical ones are: 9, 4.5, 3, 2.25, 1.8, 1.5, and the differences are always less than the count of 1. Since the Zipf distribution models large data sets, the accord for the first four values is remarkable. Quite naturally, in the very small set of our example, the correspondence for the points at the tail will not be close.

The graph of Figure 1 shows the approximation using b = 1.15.

The application of the Zipf law to the frequencies of Table 3 is consistent with the property that has been observed for all languages [25][26]. Random choice from a set (one's vocabulary) and the principle of least effort have been proposed to explain this phenomenon [27] but there is no general agreement on the underlying process [29].

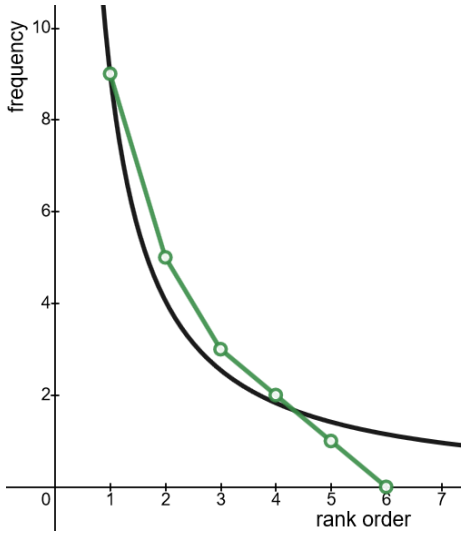

Figure 6. The rank order frequencies (green) and approximating it by $9r^{-1.15}$ (black)

Since all information spaces, including those associated with cognitive processes, will have the same optimality structure, it is quite natural to see the Zipf law show up both in codon frequencies as well as word frequencies of a natural language.

Conclusions

In this paper we considered the problem of the most efficient cover for the information space based on recent results that show that *e*-dimensionality is optimum. We showed that the assignment of the codons to the amino acids is a fractal and given by an approximate Zipf distribution that is quite consistent with the *e*-dimensionality view of the genetic information space. It is remarkable that the letter frequencies of words in a language and the triplets that code for amino acids are governed by the same statistical law.

The reason why the codon assignment for different amino acids varies is because uniformity is a requirement for optimality only in a standard vector space, and it is not so in our noninteger dimensional space that are characterized by scale invariance.

It was also shown that the number of information carrying cells in this information space is 20, and this may well be the explanation for the count of amino acids that goes into the building of proteins.

REFERENCES

1. Crick, F.H.C., Barnett, L., Brenner, S., Watts-Tobin, R.J., General nature of the genetic code for proteins. Nature 192, 1227- 1232 (1961)
2. Crick F.H.C. The origin of the genetic code. J. Mol. Biol.38, 367–379 (1982)
3. Crick F.H.C., Brenner S, Klug A, Pieczenik GA. Speculation on the origin of protein synthesis. Orig. Life7, 389–397 (1976)
4. Ribas de Pouplana L., Schimmel P. Aminoacyl-tRNA synthetases: potential markers of genetic code development. Trends Biochem. Sci.26, 591–596 (2001)
5. Rodin S.N., Rodin S.A. On the origin of the genetic code: signatures of its primordial complementarity in tRNAs and aminoacyl-tRNA synthetases. Heredity 100, 341–355 (2008)
6. Chevance, F.F.V., Hughes, K.T. Case for the genetic code as a triplet of triplets. Proc Natl Acad Sci U S A. 114, 4745–4750 (2017)
7. José, M. V., Zamudio, G. S., & Morgado, E. R. A unified model of the standard genetic code. Royal Society open science, 4(3), 160908 (2017)
8. Meierhenrich, U. Amino Acids and the Asymmetry of Life. Springer (2008)
9. Berlekamp, E.R. Algebraic Coding Theory. World Scientific (2014)
10. Kak, S. Logic of representation and information. TechRxiv (2021). https://www.techrxiv.org/articles/preprint/Logic_of_Representation_and_Information/13601939
11. Kak, S. Information theory and dimensionality of space. Scientific Reports 10, 20733 (2020). https://www.nature.com/articles/s41598-020-77855-9
12. Kak, S. The base-e representation of numbers and the power law. Circuits Syst. Signal Process. 40, 490–500 (2021); https://doi.org/10.1007/s00034-020-01480-0
13. Kak, S. The intrinsic dimensionality of data. Circuits Syst. Signal Process. 40, 2599–2607 (2021); https://doi.org/10.1007/s00034-020-01583-8
14. Hood, L., Galas, D. The digital code of DNA. Nature 421, 444-448 (2003)
15. Hurst, S.L. Multiple-valued logic - Its status and its future. IEEE Trans. Computers, C-33, 1160–1179 (1984)
16. Kak, S. On ternary coding and three-valued logic. arXiv (2018); arXiv:1807.06419
17. Kak, S. Asymptotic freedom in noninteger spaces. Scientific Reports 11, 1–5 (2021). https://www.nature.com/articles/s41598-021-83002-9
18. Kak, S. The measure of space. OSU, 2021. https://www.academia.edu/49175956/The_Measure_of_Space
19. Kak, S. Information-theoretic view of the variation of the gravitational constant. (2021). TechRxiv.
20. Falconer, K.J., Fractal Geometry: Mathematical Foundations and Applications. (Wiley, 2003)
21. Kak, S. Power series models of self-similarity in social networks. Information Sciences, 376, 31-38 (2017)
22. Bunde, A., Havlin, S., Fractals in Science. Springer (2013)
23. Vicsek, T. Fluctuations and scaling in biology. Oxford University Press (2001)
24. Kak, S. Fractals with optimum information dimension. Circuits Syst. Signal Process. 40 (2021); https://link.springer.com/article/10.1007/s00034-021-01726-5

25. Zipf, G.K. Human Behavior and the Principle of Least Effort. Addison-Wesley, Reading, MA (1949)
26. Newman, M.E.J. Power laws, Pareto distributions and Zipf's law. Contemporary Physics 46, 323-351 (2005)
27. Kak, S. Variations on the Newcomb-Benford law. arXiv (2018); https://arxiv.org/abs/1806.06695
28. Lin, H.W. and Loeb, A. Zipf's law from scale-free geometry. Phys. Rev. E 93, 032306 (2016)