Genetic Code, Hamming Distance and Stochastic Matrices

Matthew He
Division of Math, Science and Technology
Nova Southeastern University

Abstract: In the last decade the field of mathematical and computational biology has expanded very rapidly. Biological research furnishes both data on and insight into the workings of biological systems. However, qualitative and quantitative modeling and simulation are still far from allowing current knowledge to be organized into a well-understood structure. In this paper we construct a square matrix based on the genetic code. The matrix is derived from the numeric values of Hamming distance on the genetic code. The 2-bits Gray code {00, 01, 10, 11} was assigned corresponding to the genetic bases {C, A, G, U}. These square matrices are stochastic in nature and demonstrate fractal similarity properties. Furthermore the powers of these matrices are also stochastic. They resemble the similar properties to the original stochastic matrices.

1. Introduction

A mathematical view of genetic code is a map

$$\mathbf{g}: C \to A,$$

where $\mathbf{C} = \{(x_1 x_2 x_3): x_i \in \mathbf{R} = \{A, C, G, U\}\}$ = the set of codons and $\mathbf{A}$ = {Ala, Arg, Asp, …, Val, UAA, UAG, UGA}= the set of amino acids and termination codon. The inheritable information is encoded by the texts from three-alphabetic words - *triplets* or *codonums* compounded on the basis of the alphabet consisted of four characters being the nitrogen bases: A (adenine), C (cytosine), G (guanine), T (thiamine). The number of variants of location of 64 triplets in octet tables is equal 64! or 10^{89} approximately. It's unimaginably huge number. The following three genetic code tables have appeared in various literature due to its symmetrical structure and biochemical properties.

In the first table the codons are arranged in an 8×8 square pattern along with their number of hydrogen bonds. In this square both the row and column numbers are labeled in the standard Gray code, e.g., 000, 001, 011, 010, 110, 111, 101, 100, and each element of the table is listed by a 6-bit representation. It has been found that the amino acids are formed from contiguous groups of codons, e.g., proline: CCC, CCU, CCA, CCG; glutamine: CAA, CAG; leucine: CUU, CUC, CUG, CUA, UUA, UUG; etc. [7].

CCC	**CCU**	**CUU**	**CUC**	**UUC**	**UUU**	**UCU**	**UCC**
CCA	**CCG**	**CUG**	**CUA**	**UUA**	**UUG**	**UCG**	**UCA**
CAA	**CAG**	**CGG**	**CGA**	**UGA**	**UGG**	**UAG**	**UAA**
CAC	**CAU**	**CGU**	**CGC**	**UGC**	**UGU**	**UAU**	**UAC**
ACC	**AAU**	**AGU**	**AGC**	**GGC**	**GGU**	**GAU**	**GAC**
AAA	**AAG**	**AGG**	**AGA**	**GGA**	**CGG**	**GAG**	**GAA**
ACA	**ACG**	**AUG**	**AUA**	**GUA**	**GUG**	**GCG**	**GCA**
ACC	**ACU**	**AUU**	**AUC**	**GUC**	**GUU**	**GCU**	**GCC**

Table 1 (Gray code based)

The second table is called bi-periodic table introduced in [ ]. The table 2, in accordance with determined biochemical data, is unique variant, which reveals natural ordering in set of triplets and demonstrates new structural properties of genetic system.

CCC	CCA	CAC	CAA	ACC	ACA	AAC	AAA
CCU	CCG	CAU	CAG	ACU	ACG	AAU	AAG
CUC	CUA	CGC	CGA	AUC	AUA	AGC	AGA
UCC	UCA	UAC	UAA	GCC	GCA	GAC	GAA
CUU	CUG	CGU	CGG	AUU	AUG	AGU	AGG
UCU	UCG	UAU	UAG	GCU	GCG	GAU	GAG
UUC	UUA	UGC	UGA	GUC	GUA	GGC	GGA
UUU	UUG	UGU	UGG	GUU	GUG	GGU	GGG

Table 2 (Biperiodic Table)

The third table is generated based on a 4-ary tree. The symmetrical structures of the genetic code were recently studied by the author (He, 2003).

AAA	ACA	AGA	AUA	CAA	CCA	CGA	CUA
AAC	ACC	AGC	AUC	CAC	CCC	CGC	CUC
AAG	ACG	AGG	AUG	CAG	CCG	CGG	CUG
AAU	ACU	AGU	AUU	CAU	CCU	CGU	CUU
GAA	GCA	GGA	GUA	UAA	UCA	UGA	UUA
GAC	GCC	GGC	GUC	UAC	UCC	UGC	UUC
GAG	GCG	GGG	GUG	UAG	UCG	UGG	UUG
GAU	GCU	GGU	GUU	UAU	UCU	UGU	UUU

Three attribute-based mappings were introduced in [ ] to connect the genetic code with the matrices. The resulting square matrices were shown to be stochastic. In this paper we use the Gray code to connect the genetic code and generate the numeric matrices based on the Table 1, Table 2 and Table 3 .

2. Gray Code and Hamming Distance

A binary code in which consecutive decimal numbers are represented by binary expressions that differ in the state of one, and only one, one bit. A Gray code was used in a telegraph demonstrated by French engineer Émile Baudot in 1878. The codes were first patented by Frank Gray, a Bell Labs researcher, in 1953.

One way to construct a Gray code for n bits is to take a Gray code for n-1 bits with each code prefixed by 0 (for the first half of the code) and append the n-1 Gray code reversed with each code prefixed by 1 (for the second half). This is called a "binary-reflected Gray code". Here is an example of creating a 3-bit Gray code from a 2-bit Gray code.

00	**01**	**11**	**10**					**A Gray code for 2 bits**
000	001	011	010					the 2-bit code with "0" prefixes
				10	11	01	00	the 2-bit code in reverse order
				110	111	101	100	the reversed code with "1" prefixes
000	**001**	**011**	**010**	**110**	**111**	**101**	**100**	**A Gray code for 3 bits**

The Hamming distance H is defined only for strings of the same length. For two strings s and t, H(s,t) is the number of places in which the two string differ, i.e., have different characters. More formally, the distance between two strings A and B is $\Sigma| A_i - B_i |$. E.g., 0101 and 0110 has a Hamming distance of two whereas "Butter" and "ladder" are four characters apart. The Hamming distance between 2143896 and 2233796 is three, and between "toned" and "roses" it is also three.

This distance is applicable to encoded information, and is a particularly simple metric of comparison, often more useful than the *city-block distance* (the sum of absolute values of distances along the coordinate axes) or Euclidean distance (the square root of the sum of squares of the distances along the coordinate axes).

There is a natural way to relate the genetic codons to Gray code. A 2-bit binary Gray code has four possible bases {00, 01, 11, 10}. We use the following assignments:

RNA BASES	Binary 2-Bit Gray Code
C	**0** **0**
A	**0** **1**
G	**1** **1**
U	**1** **0**

For example CUG → 011 and GAC → 100. GAC is called the anti-codon of CUG since
001 110

1 of the codon is replaced by 0 and 0 by 1 of the codon to get the anti-codon. Notice that the upper and lower bit strings of both the codon and anti-codon differ in a single bit, e.g., they have a Hamming distance of 1. It has been found that the amino acids are formed from contiguous groups of codons, e.g., proline: CCC, CCU, CCA, CCG; glutamine: CAA, CAG; leucine: CUU, CUC, CUG, CUA, UUA, UUG; etc. [7]. Apparently Gray code arises in genetics as a means of minimizing the "cliffs" or mismatches between the digits encoding adjacent bases and therefore the degree of mutation or differences between nearby chromosome segments. The requirement in an encoding scheme is that changing one bit in the segment of the chromosome should cause that segment to map to an element which is adjacent to the premutated element.

3. Hamming Distance and Stochastic Matrix

In this section, we first give the Gray code assignments to each table from the Section 1 and then compute the corresponding Hamming distance corresponding to each codon. A set of three 8x8 square matrices for Tables 1, 2 and 3 will be produced. These three matrices illustrate different mathematical structure. It turns out that the matrix generated from the Table 2 has optimal stochastic symmetry.

	000	**001**	**011**	**010**	**110**	**111**	**101**	**100**
000	**000** **000**	**001** **000**	**011** **000**	**010** **000**	**110** **000**	**111** **000**	**101** **000**	**100** **000**
001	**000** **001**	**001** **001**	**011** **001**	**010** **001**	**110** **001**	**111** **001**	**101** **001**	**100** **001**
011	**000** **011**	**001** **011**	**011** **011**	**010** **011**	**110** **011**	**111** **011**	**101** **011**	**100** **011**
010	**000** **010**	**001** **010**	**011** **010**	**010** **010**	**110** **010**	**111** **010**	**101** **010**	**100** **010**
110	**000** **110**	**001** **110**	**011** **110**	**010** **110**	**110** **110**	**111** **110**	**101** **110**	**100** **110**
111	**000** **111**	**001** **111**	**011** **111**	**010** **111**	**110** **111**	**111** **111**	**101** **111**	**100** **111**
101	**000** **101**	**001** **101**	**011** **101**	**010** **101**	**110** **101**	**111** **101**	**101** **101**	**100** **101**
100	**000** **100**	**001** **100**	**011** **100**	**010** **100**	**110** **100**	**111** **100**	**101** **100**	**100** **100**

Table 1: Gray code based

	000	001	010	011	100	101	110	111
000	000 000	000 001	000 010	000 011	000 100	000 101	000 110	000 111
001	001 000	001 001	001 010	001 011	001 100	001 101	001 110	001 111
010	010 000	010 001	010 010	010 011	010 100	010 101	010 110	010 111
100	100 000	100 001	100 010	100 011	100 100	100 101	100 110	100 111
011	011 000	011 001	011 010	011 011	011 100	011 101	011 110	011 111
101	101 000	101 001	101 010	101 011	101 100	101 101	101 110	101 111
110	110 000	110 001	110 010	110 011	110 100	110 101	110 110	110 111
111	111 000	111 001	111 010	111 011	111 100	111 101	111 110	111 111

Table 2: Biperiodic Table Based

000 111	000 101	010 111	010 101	000 011	000 001	010 011	010 001
000 110	000 100	010 110	010 100	000 010	000 000	010 010	010 000
001 111	001 101	011 111	011 101	001 011	001 001	011 011	011 001
001 110	001 100	011 110	011 100	001 010	001 000	011 010	011 000
100 111	100 101	110 111	110 101	100 011	100 001	110 011	110 001
100 110	100 100	110 110	110 100	100 010	100 000	110 010	110 000
101 111	101 101	111 111	111 101	101 011	010 000	111 011	111 001
101 110	101 100	111 110	111 100	101 010	101 000	111 010	111 000

Table 3: 4-ary Tree Based

Next we compute the Hamming distance for each codon and list all three matrices.

0	**1**	**2**	**1**	**2**	**3**	**2**	**1**
1	**0**	**1**	**2**	**3**	**2**	**1**	**2**
2	**1**	**0**	**1**	**2**	**1**	**2**	**3**
1	**2**	**1**	**0**	**1**	**2**	**3**	**2**
2	**3**	**2**	**1**	**0**	**1**	**2**	**1**
3	**2**	**1**	**2**	**1**	**0**	**1**	**2**
2	**1**	**2**	**3**	**2**	**1**	**0**	**1**
1	**2**	**3**	**2**	**1**	**2**	**1**	**0**

Hamming distance matrix of gray code table Doubly stochastic 2/4/6/8)

0	**1**	**1**	**2**	**1**	**2**	**2**	**3**
1	**0**	**2**	**1**	**2**	**1**	**3**	**2**
1	**2**	**0**	**1**	**2**	**3**	**1**	**2**
1	**2**	**2**	**3**	**0**	**1**	**1**	**2**
2	**1**	**1**	**0**	**3**	**2**	**2**	**1**
2	**1**	**3**	**2**	**1**	**0**	**2**	**1**
2	**3**	**1**	**2**	**1**	**2**	**0**	**1**
3	**2**	**2**	**1**	**2**	**1**	**1**	**0**

Hamming distance matrix of biperiodical table (Doubly stochastic 2/4/6/8)

3	**2**	**2**	**3**	**2**	**1**	**1**	**2**
2	1	1	2	1	0	0	**1**
2	1	1	2	1	0	0	**1**
3	2	2	3	2	1	1	**2**
2	1	1	2	3	2	2	**3**
1	0	0	1	2	1	1	**2**
1	0	0	1	2	1	1	**2**
2	**1**	**1**	**2**	**3**	**2**	**2**	**3**

Hamming distance matrix of 4-ary table

It’s easy to see that the first two matrices are doubly stochastic. The third matrix is not stochastic.

Hamming Distance and frequency of genetic code

Hamming Distance	**Genetic Code Frequency**	**Sum of Distance**

0	8	0
1	24	24
2	24	48
3	8	24
Total	**64**	**96**

Matrix Configurations of Genetic Code (64 Codons)

Matrix	**Dimension**
G [1,64] = G [2^0, 2^6]	1 x 64 = 2^0x2^6
G [2,32 = G [2^1, 2^5]	2 x 32 = 2^1x2^5
G [4,16] = G [2^2, 2^4]	4 x 16 = 2^2x2^4
G [8,8] = G [2^3, 2^3]	8 x 8 = 2^3x2^3

Stochastic Matrix	**Common Row Sums**	**Common Column Sums**
G [1,64] = G [2^0, 2^6]	N/A	96
G [2,32] = G [2^1, 2^5]	3	48
G [4,16] = G [2^2, 2^4]	6	24
G [8,8] = G [2^3, 2^3]	12	12

G [8, 8] is a doubly stochastic matrix.

Matrix configurations with defined common row and column sum

G [1, 64]

0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 3 3 3 3 3 3 3 3

G [2, 32]

0	0	0	0	0	0	0	0	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
3	3	3	3	3	3	3	3	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

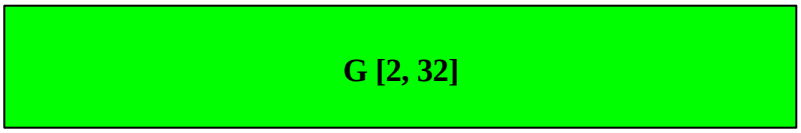
G [4, 16]

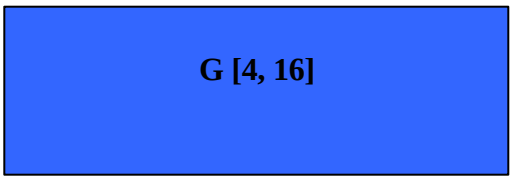
0	1	2	1	2	1	2	3	0	1	2	1	2	1	2	3
3	2	1	2	1	2	1	0	3	2	1	2	1	2	1	0
0	1	2	1	2	1	2	3	0	1	2	1	2	1	2	3
3	2	1	2	1	2	1	0	3	2	1	2	1	2	1	0

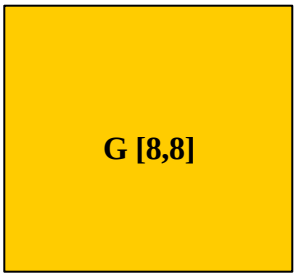
G [8, 8]

0	1	1	2	1	2	2	3
1	0	2	1	2	1	3	2
1	2	0	1	2	3	1	2
1	2	2	3	0	1	1	2
2	1	1	0	3	2	2	1
2	1	3	2	1	0	2	1
2	3	1	2	1	2	0	1
3	2	2	1	2	1	1	0

We next illustrate these four types of matrices by rectangles. One can easily see that the square shape has the maximum area if the total parameter of rectangles is a constant.







0	1	1	1	1	2	1	2	2	2	2	3
1	0	1	1	1	2	1	2	2	2	3	2
1	1	0	1	1	2	1	2	2	3	2	2
1	1	1	0	2	1	2	1	3	2	2	2
1	1	1	2	0	1	2	3	1	2	2	2
1	1	1	2	2	3	0	1	1	2	2	2

2	2	2	1	1	0	3	2	2	1	1	1
2	2	2	1	3	2	1	0	2	1	1	1
2	2	2	3	1	2	1	2	0	1	1	1
2	2	3	2	2	1	2	1	1	0	1	1
2	3	2	2	2	1	2	1	1	1	0	1
3	2	2	2	2	1	2	1	1	1	1	0

Levenshtein Distance

The Levenshtein (or *edit*) distance is more sophisticated. It's defined for strings of arbitrary length. It counts the differences between two strings, where we would count a difference not only when strings have different characters but also when one has a character whereas the other does not. The formal definition follows.

For a string s, let s(i) stand for its i-th character. For two characters a and b, define

r(a, b) = 0 if a = b. Let r(a, b) = 1, otherwise.

Assume we are given two strings s and t of length n and m, respectively. We are going to fill an (n+1)x(m+1) array d with integers such that the low right corner element d(n+1, m+1) will furnish the required values of the Levenshtein distance L(s,t).

The definition of entries of d is recursive. First set d(i,0)=i, i=0,1,...,n, and d(0,j)=j, j=0,1,...,m. For other pairs i,j use

d(i,j) = min(d(i-1, j)+1, d(i, j-1)+1, d(i-1, j-1) + r(s(i), t(j)))

	Second nucleotide				
	U	C	A	G	
U	UUU (3) 111 000	UCU (2) 101 000	UAU (3) 101 010	UGU (2) 111 010	U
	UUC (2) 110 000	UCC (1) 100 000	UAC (2) 100 010	UGC (1) 110 010	C
	UUA (3) 110 001	UCA (2) 100 001	UAA (3) 100 011	UGA (2) 110 011	A
	UUG (2) 111 001	UCG (1) 101 001	UAG (2) 101 011	UGG (1) 111 011	G
C	CUU (2) 011 000	CCU (1) 001 000	CAU (2) 001 010	CGU (1) 011 010	U
	CUC (1) 010 000	CCC (0) 000 000	CAC (1) 000 010	CGC (0) 010 010	C
	CUA (2) 010 001	CCA (1) 000 001	CAA (2) 000 011	CGA (1) 010 011	A
	CUG (1) 011 001	CCG (0) 001 001	CAG (1) 001 011	CGG (0) 011 011	G
A	AUU (3) 011 100	ACU (2) 001 100	AAU (3) 001 110	AGU (2) 011 110	U
	AUC (2) 010 100	ACC (1) 000 100	AAC (2) 000 110	AGC (1) 010 110	C
	AUA (3) 010 101	ACA (2) 000 101	AAA (3) 000 111	AGA (2) 010 111	A
	AUG (2) 011 101	ACG (1) 001 101	AAG (2) 001 111	AGG (1) 011 111	G
G	GUU (2) 111 100	GCU (1) 101 100	GAU (2) 101 110	GGU (1) 111 110	U
	GUC (1) 110 100	GCC (0) 100 100	GAC (1) 100 110	GGC (0) 110 110	C
	GUA (2) 110 101	GCA (1) 100 101	GAA (2) 100 111	GGA (1) 110 111	A
	GUG (1) 111 101	GCG (0) 101 101	GAG (1) 101 111	GGG (0) 111 111	G

Reference

He, M. (2003), Symmetry in Structure of Genetic Code. *Proceedings of the Third All-Russian Interdisciplinary Scientific Conference "Ethics and the Science of Future. Unity in Diversity"*, February 12-14, Moscow.

He, M. (2003), Genetic Code, Attributive Mappings and Stochastic Matrices, submitted for publication.

Petoukhov, S.V. (2001), *The Bi-periodic Table of Genetic Code and Number of Protons*, Foreword of K. V. Frolov, Moscow, 258 (in Russian).

Petoukhov, S. V. (2002), Binary sub-alphabets of genetic language and problem of unification bases of biological languages, *IX International Conference "Mathematics, computer, education"*, Russia, Dubna, January 28-31, 191 (in Russian).

Romanovsky, V. (1931), Sur les zeros des matrices stocastiques, C. R. Acad. Sci. Paris 192, 266-269. [Zbl. 1 (1932) 055]

Walter, K.: *Tao of Chaos: Merging East and West*, Kairos Center, Austin, 1994.