Bulletin of Mathematical Biology (2004) **66**, 1405–1421
doi:10.1016/j.bulm.2004.01.002

Genetic Code, Hamming Distance and Stochastic Matrices

MATTHEW X. HE*
Division of Math, Science and Technology,
Nova Southeastern University,
3301 College Avenue,
Fort Lauderdale, FL 33314,
USA
E-mail: hem@nova.edu

SERGEI V. PETOUKHOV
Department of Biomechanics, Mechanical Engineering Research Institute,
Russian Academy of Sciences,
Moscow, 101830,
Russia

PAOLO E. RICCI
Dipartimento di Matematica “Guido Castelnuovo”,
Università degli Studi di Roma,
“La Sapienza”,
Rome,
Italy

In this paper we use the Gray code representation of the genetic code C = 00, U = 10, G = 11 and A = 01 (C pairs with G, A pairs with U) to generate a sequence of genetic code-based matrices. In connection with these code-based matrices, we use the Hamming distance to generate a sequence of numerical matrices. We then further investigate the properties of the numerical matrices and show that they are doubly stochastic and symmetric. We determine the frequency distributions of the Hamming distances, building blocks of the matrices, decomposition and iterations of matrices. We present an explicit decomposition formula for the genetic code-based matrix in terms of permutation matrices, which provides a hypercube representation of the genetic code. It is also observed that there is a Hamiltonian cycle in a genetic code-based hypercube.

1. INTRODUCTION

The universal genetic code is the mapping of nucleic acids into polypeptides that is employed in every organism, organelle and virus with some minor variations.

*Author to whom correspondence should be addressed.

Table 1. Biperiodic table of genetic code by Petoukhov.

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

A mathematical view of genetic code is a map

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

where $\mathbf{C} = \{(x_1 x_2 x_3) \colon x_i \in \mathbf{R} = \{\mathrm{A, C, G, U}\}\} =$ the set of codons and $\mathbf{A} =$ {Ala, Arg, Asp, . . . , Val, UAA, UAG, UGA} = the set of amino acids and termination codons. The inheritable information is encoded by the texts from three-alphabetic words—*triplets* compounded on the basis of the alphabet consisting 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 to 64! or 10^{89} approximately. It is an unimaginably huge number. There have been many attempts to give a formal characterization of the particular structure of the code which would also have a justification from physicochemical and/or evolutionary points of view (Knight *et al*., 1999). A Gray code representation of the genetic code was proposed by Swanson in Swanson (1984). A representation of the genetic code as a 6-dimensional Boolean hypercube was proposed in Jimenéz-Montaño *et al*. (1994). A topological approach was also applied in Yang (2003) to rearranging the Hamiltonian-type graph of the codon map into a polyhedron model. In Štambuk (2000), universal metric properties of the genetic code were defined by means of the nucleotide base representation on the square with vertices U or T = 00, C = 01, G = 10 and A = 11. It was shown that this notation defines the Cantor set and Smale horseshoe map representation of the genetic code. Recently Petoukhov discovered the 'Biperiodic table of genetic code' (Petoukhov, 1999, 2001, 2002) shown in Table 1. This table demonstrates a great symmetrical structure and has led to many discoveries. The stochastic characteristic of the biperiodic table and symmetries in structure of genetic code were recently investigated in He (2003a,b,c) by using three fundamental attributive mappings.

In this paper we use the Gray code representation of the genetic code C = 00, U = 10, G = 11 and A = 01 to generate a sequence of genetic code-based matrices. In connection with these code-based matrices, we use the Hamming distance to generate a sequence of *numerical matrices*. We investigate the properties of the

Table 2. Gray code generation.

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**

Table 3. Gray code.

$\mathbf{G}_n$	Gray code sequence $\mathbf{s}_n$	Number of $\mathbf{G}_n$
$\mathbf{G}_1$	0	
	1	$2^1 = 2$
$\mathbf{G}_2$	00 01	
	11 10	$2^2 = 4$
$\mathbf{G}_3$	000 001 011 010	
	110 111 101 100	$2^3 = 8$
.	...	.
.	...	.
$\mathbf{G}_n$	$000\cdots0\,000\cdots1\,000\cdots11\cdots010\cdots0$	
	$111\cdots0\,111\cdots1\,111\cdots00\cdots100\cdots0$	2^n

numerical matrices and show that they are doubly stochastic and symmetric. We determine the frequency distributions of the Hamming distances, building blocks of the matrices, decomposition and iterations of matrices. We shall present an explicit decomposition formula for the genetic code-based matrix in terms of permutation matrices, which provide a hypercube representation of the genetic code.

2. Gray Code, Hamming Distance and Stochastic Matrices

A Gray code was used in a telegraph demonstrated by French engineer Émile Baudot in 1878. The codes were first patented by Frank Gray in 1953. The Gray code is a binary code in which consecutive decimal numbers are represented by binary expressions that differ in the state of one, and only one, bit. Gray codes have been extensively studied in other contexts. For example, Gray codes have been used in converting analog information to digital form. Here we review briefly how to construct a Gray code for each positive integer n. 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 (Table 2).

We next list the sequences of the Gray codes denoted by G_n in Table 3. It is easy to see that every n-bit string appears somewhere in the sequence; sequences

Table 4. RNA bases and Gray code.

RNA bases	Binary 2-bit Gray code
C	0 0
A	0 1
G	1 1
U	1 0

s_i, s_{i+1} differ in exactly one bit, $i = 1, 2, \ldots, 2^n - 1$; ${s_2}^n$ and s_1 differ in exactly one bit. This proves that the n-cube has a Hamiltonian cycle for every positive integer $n \geq 2$, for example, $s_1, s_2, \ldots, {s_2}^n, s_1$ is a Hamiltonian cycle. There is a natural way to relate the genetic codon to Gray code. A 2-bit binary Gray code has four possible bases $\{00, 01, 11, 10\}$. We use the assignments given in Table 4:

For example CUG → $\begin{matrix}011\\001\end{matrix}$ and GAC → $\begin{matrix}100\\110\end{matrix}$. GAC is called the anti-codon of CUG since 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. The Gray code arises in genetics as a means of minimizing the 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 pre-mutated element.

The Hamming distance D is defined for strings of the same length. For two strings s and t, $D(s, t)$ is the number of places in which the two strings differ, i.e., have different characters. More formally, the distance between two strings A and B is $\Sigma|A_i - B_i|$. For example, 0101 and 0110 has a Hamming distance of two whereas 'Butter' and 'ladder' are four characters apart. The Hamming distance between 2143 896 and 2233 796 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.

Next we formalize our algorithm to generate the Hamming distance-based matrices corresponding to genetic code-based matrices. Let n be the length of strings (binary string or DNA/RNA strings). We present our constructions for $n = 1, 2$, and 3. The general result will be summarized as a theorem for any positive integer n.

For $n = 1$, the Gray code $\mathbf{G_1} = \{0, 1\}$. We arrange the G_1 in a 2-dimensional table (row/column) and form the table entry by stacking the column code on top of

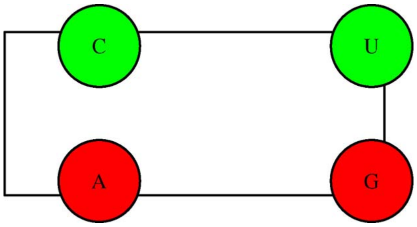

Figure 1. RNA single base.

the row code as below. Denote this matrix by $\mathbf{H_{21}}$. This is a 2 × 2 matrix generated by G_1.

$\mathbf{G_1}$	**0**	**1**
0	0 0	1 0
1	0 1	1 1

The corresponding genetic code-based matrix with a single base is denoted by $\mathbf{C_{21}}$ as below.

C	U
A	G

We next compute the Hamming distance of each entry of matrix H_{21}. The resulting matrix is denoted by $\mathbf{D_{21}}$,

0	1
1	0

This matrix $\mathbf{D_{21}}$ has Hamming distances 0's and 1's. The frequencies of the 0's and 1's are 2 and 2, respectively. The total sum of the matrix is 2. The common row/column sum is 1. The Hamming distance between any two horizontal and vertical neighboring entries is 1. We illustrate this property by a diagram (Fig. 1).

For $n = 2$, the Gray code $\mathbf{G_2} = \{00, 01, 11, 10\}$. We arrange the G_2 in a 2-dimensional table and form the table entry by stacking the column code on top of the row code as below. Denote this matrix by $\mathbf{H_{42}}$. This is a 4 × 4 matrix generated by G_2.

$\mathbf{G_2}$	**00**	**01**	**11**	**10**
00	00 00	01 00	11 00	10 00
01	00 01	01 01	11 01	10 01
11	00 11	01 11	11 11	10 11
10	00 10	01 10	11 10	10 10

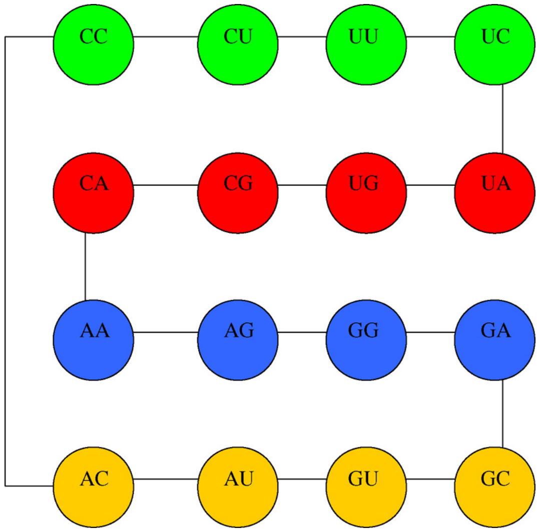

Figure 2. RNA duplets.

The corresponding genetic code-based matrix with a single base is denoted by $\mathbf{C_{42}}$ as below.

CC	CU	UU	UC
CA	CG	UG	UA
AA	AG	GG	GA
AC	AU	GU	GC

We next compute the Hamming distance of each entry of the matrix H_{42}. The resulting matrix is denoted by $\mathbf{D_{42}}$,

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

or (blocked differently),

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

We note that the matrix $\mathbf{D_{21}}$ is centrally embedded inside $\mathbf{D_{42}}$ and the matrix $\mathbf{D_{42}}$ has two 2 × 2 matrices building blocks denoted by B_{21} and B_{22}. The matrix $\mathbf{D_{42}}$ may be written as

B_{21}	B_{22}
B_{22}	B_{21}

The frequencies of matrix building blocks B_{21} and B_{22} are 2 and 2, respectively. This matrix $\mathbf{D_{42}}$ has Hamming distances 0's, 1's, and 2's. The frequencies of the 0's, 1's and 2's are 4, 8, and 4, respectively. The total sum of the matrix is 16. The common row/column sum is 4. The Hamming distance between any two horizontal and vertical neighboring entries is 1. We illustrate this property by a diagram (Fig. 2).

For $n = 3$, the Gray code $\mathbf{G_3} = \{000, 001, 011, 010, 110, 111, 101, 100\}$. The matrix $\mathbf{H_{82}}$ is a 8×8 square matrix.

G_3	**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

The corresponding genetic code-based matrix with a single base is denoted by $\mathbf{C_{83}}$ as below.

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
AAC	AAU	AGU	AGC	GGC	GGU	GAU	GAC
AAA	AAG	AGG	AGA	GGA	GGG	GAG	GAA
ACA	ACG	AUG	AUA	GUA	GUG	GCG	GCA
ACC	ACU	AUU	AUC	GUC	GUU	GCU	GCC

The matrix $\mathbf{C_{83}}$ is another representation of universal genetic code. It contains all 64 codons. 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. (Swanson, 1984). The Hamming distance-based matrix $\mathbf{D_{83}}$ is shown below:

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**

or

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**

This matrix $\mathbf{D_{83}}$ has Hamming distances 0's, 1's, 2's and 3's. The frequencies of the 0's, 1's, 2's and 3's are 8, 24, 24, and 8, respectively. The total sum of the matrix $\mathbf{D_{83}}$ is 96. The common row/column sum is 12. We also note that the matrices $\mathbf{D_{21}}$ and $\mathbf{D_{42}}$ are centrally embedded inside $\mathbf{D_{83}}$ and the matrix $\mathbf{D_{83}}$ has three 2×2 matrices building blocks B_{21}, B_{22} and B_{23}. The matrix $\mathbf{D_{83}}$ may also be written as

B_{21}	B_{22}	B_{23}	B_{22}
B_{22}	B_{21}	B_{22}	B_{23}
B_{23}	B_{22}	B_{21}	B_{22}
B_{22}	B_{23}	B_{22}	B_{21}

The frequencies of matrix building blocks B_{21}, B_{22} and B_{23} are 4, 8, and 4, respectively. The distribution of the genetic code and frequencies of Hamming distances are readily observed (Table 5).

The Hamming distance between any two horizontal and vertical neighboring entries is 1. We illustrate this property by a diagram (Fig. 3).

For general positive integer n, we have the following results.

THEOREM A. *Let n be the length of binary or DNA/RNA strings and G_n be the n-bit Gray code. Then*

Table 5. Genetic code frequency.

Distance	Codons	Frequency
0	CCC CCG CGG CGC GGC CGG GCG GCC	8
1	ACC ACG AGC AGG CAC CAG CCA CGA CCU CGU CUC CUG GAC GAG GCA GGA GCU GGU GUC GUG UCC UGC UCG UGG	24
2	AAC AAG ACA ACU AGA AGU AUC AUG CAA CAU CUA CUU GAA GAU GUA GUU UAC UAG UCA UCU UGA UGU UUC UUG	24
3	AAA AAU AUA AUU UAA UAU UUA UUU	8

(i) *The genetic code-based matrix* $\mathrm{C}_{2}{}^{n}{}_{n}$ *is a* $2^n \times 2^n$ *matrix with RNA bases of length* n*. Each two neighboring entries of the genetic code both from vertical and horizontal direction differs exactly one base.*

(ii) *The Hamming distance-based matrix* $\mathrm{D}_{2}{}^{n}{}_{n}$ *is also a* $2^n \times 2^n$ *matrix with Hamming distances of* $0, 1, 2, \ldots, n$*. The common row/column sum of the matrix* $\mathrm{D}_{2}{}^{n}{}_{n}$ *equals* $n2^{n-1}$ *and the total summation of the entries of matrix* $\mathrm{D}_{2}{}^{n}{}_{n}$ *is* $n2^{2n-1}$.

(iii) *The matrix* $\mathrm{D}_{2}{}^{n}{}_{n}$ *is doubly stochastic and symmetric.*

(iv) *The frequency distributions denoted by* $f_{nk}(n = 2, 3, \ldots, k = 1, 2, \ldots)$ *of Hamming distances of* $0, 1, 2, \ldots, n$ *are shown in* Table 6 *for* $n = 1, 2, 3, 4, 5,$ *and* 6.

Table 6. Frequency distribution of Hamming distances.

n	Hamming distances	Frequency distributions	Frequency notation
1	0 1	2 2	f_{21} f_{22}
2	0 1 2	4 8 4	f_{31} f_{32} f_{33}
3	0 1 2 3	8 24 24 8	f_{41} f_{42} f_{43} f_{44}
4	0 1 2 3 4	16 64 96 64 16	f_{51} f_{52} f_{53} f_{54} f_{55}
5	0 1 2 3 4 5	32 160 320 320 160 32	f_{61} f_{62} f_{63} f_{64} f_{65} f_{66}
6	0 1 2 3 4 5 6	64 384 960 1280 960 384 64	f_{71} f_{72} f_{73} f_{74} f_{75} f_{76} f_{77}

Or

n	Hamming distances	Frequency distributions		Frequency notation
1	0 1	2	1 1	f_{21} f_{22}
2	0 1 2	4	1 2 1	f_{31} f_{32} f_{33}
3	0 1 2 3	8	1 3 3 1	f_{41} f_{42} f_{43} f_{44}
4	0 1 2 3 4	16	1 4 6 4 1	f_{51} f_{52} f_{53} f_{54} f_{55}
5	0 1 2 3 4 5	32	1 5 10 10 5 1	f_{61} f_{62} f_{63} f_{64} f_{65} f_{66}
6	0 1 2 3 4 5 6	64	1 6 15 20 15 6 1	f_{71} f_{72} f_{73} f_{74} f_{75} f_{76} f_{77}

The general relationships of the frequencies are determined by a recurrence formula:

$$f_{21} = 2, \qquad f_{22} = 2,$$

$$f_{nk} = 2(f_{(n-1)(k-1)} + f_{(n-1)k}).$$

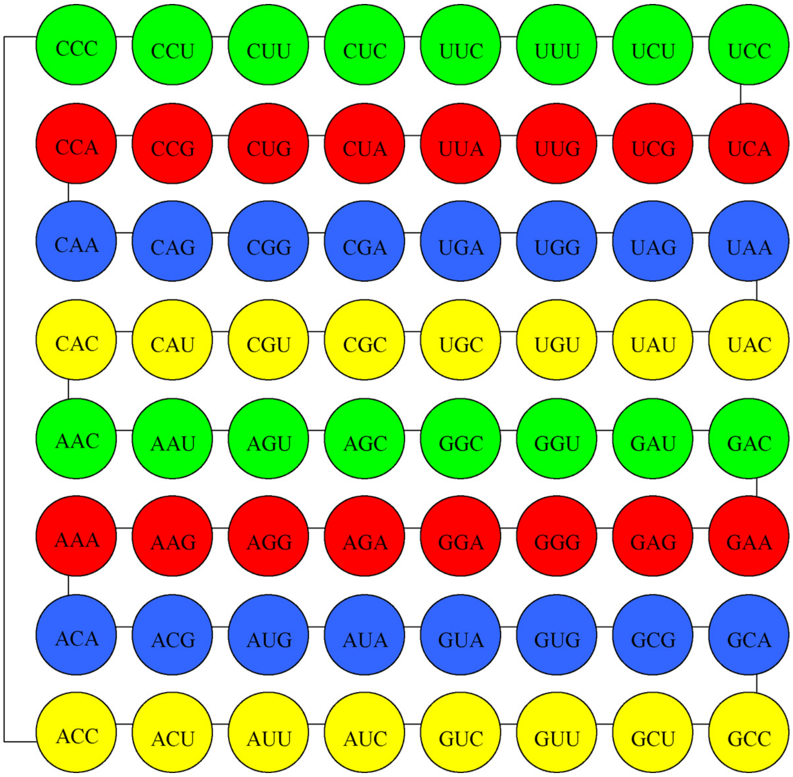

Figure 3. RNA triplets.

The frequency distribution of the Hamming distances is the ***Pascal triangle*** *with a multiple of* 2^n. *The solution to this recurrence relation is*

$$f_{nk} = 2^n \mathrm{C}(n, k), \qquad k = 1, 2, \ldots, n.$$

(v) *The matrix* $\mathrm{D}_{2^n n}$ *consists of* $(n-1)$ 2×2 *matrix building blocks* $\mathrm{B}_{21}, \mathrm{B}_{22}, \ldots, \mathrm{B}_{2(n-1)}$. *The previous matrix* $\mathrm{D}_{2^{n-1}(n-1)}$ *is centrally embedded inside the next matrix* $\mathrm{D}_{2^n n}$. *The frequencies of matrix building blocks* $\mathrm{B}_{21}, \mathrm{B}_{22}, \ldots, \mathrm{B}_{2(n-1)}$ *are* $f_{(n-1)1}, f_{(n-1)2}, \ldots, f_{(n-1)(n-1)}$, *respectively.*

In the next section we illustrate the stochastic and hypercube structure of the genetic code-based matrix $\mathrm{C}_{2^n n}$ via the structure of matrix of $\mathrm{D}_{2^n n}$.

3. Stochastic Matrices and Hypercube Structure of the Genetic Code

As we have noted, the matrix $\mathrm{D}_{2^n n}$ is a symmetric and doubly stochastic matrix. For its simplicity, we consider the case when $n = 3$, i.e., D_{83}, the entry of the matrix is a RNA codon. We first recall some basic definition of the stochastic matrix. A square matrix of $P = (p_{ij})$ is a **stochastic matrix** if all entries of the matrix are nonnegative and the sum of the elements in each row (or column) is unity or a constant. If the sum of the elements in each row and column is unity or the same, the matrix is called **doubly stochastic**. The term 'stochastic matrix' goes back at least to Romanovsky (1931). It plays a large role in the theory of discrete Markov chains. Stochastic matrices and doubly stochastic matrices have many remarkable properties. The properties of stochastic matrices are mainly spectral theoretic and are motivated by Markov chains. Doubly stochastic matrices have additional combinatorial structure. Here we list some basic properties of the matrix D_{83}.

- The matrix D_{83} is symmetric since $\mathrm{D}_{83} = {\mathrm{D}_{83}}^{\mathrm{T}}$ (the transpose of a matrix).
- The matrix D_{83} is singular since $\mathrm{Det}(\mathrm{D}_{83}) = 0$ (determinant of a matrix).
- The eigenvalues of D_{83} is $\{\lambda_1, \lambda_2, \dots, \lambda_8\} = \{-4, -4, -4, 0, 0, 0, 12\}$.
- The eigenvectors are $\{0, -1, -1, 0, 1, 0, 0, 1\}$, $\{0, 1, 0, -1, -1, 0, 1, 0\}$, $\{-1, -1, 0, 0, 1, 1, 0, 0\}$, $\{0, -1, 1, 0, -1, 0, 0, 0\}$, $\{1, -2, 1, 0, -1, 0, 1, 0\}$, $\{1, -1, 0, 0, -1, 1, 0, 0\}$, $\{-1, 1, -1, 1, 0, 0, 0, 0\}$, $\{1, 1, 1, 1, 1, 1, 1, 1\}$. Furthermore these eight vectors are linearly independent. They form a basis for a vector space of dimension of 8.
- The trace of matrix $\mathrm{D}_{83} =$ sum of eigenvalues $= 0 + 0 + 0 + 0 - 4 - 4 - 4 + 12 = 0$.

3.1. ***Decomposition of the genetic matrix.*** Since the matrix D_{83} is doubly stochastic, the matrix D_{83} can be decomposed as a convex combination of finitely many permutation matrices [14]; that is,

$$\mathrm{D}_{83} = a_1\mathrm{P}_1 + a_2\mathrm{P}_2 + \cdots + a_8\mathrm{P}_8,$$

where $\mathrm{P}_1, \mathrm{P}_2, \dots, \mathrm{P}_8$ are permutation matrices and $0 = a_1, a_2, \dots, a_8 = 12$, $a_1 + a_2 + \cdots + a_8 = 12$. A permutation matrix can be obtained from an identity matrix by permuting its rows and columns. Explicitly we have the following theorem.

Theorem B. $\mathrm{D}_{83} = 0\mathrm{P}_1 + 1(\mathrm{P}_2 + \mathrm{P}_3 + \mathrm{P}_4) + 2(\mathrm{P}_5 + \mathrm{P}_6 + \mathrm{P}_7) + 3\mathrm{P}_8$, *where*

$$\mathrm{P}_1 = \begin{array}{|c|c|c|c|c|c|c|c|}
\hline
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \hline
0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ \hline
0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ \hline
0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ \hline
0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ \hline
0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ \hline
0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ \hline
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ \hline
\end{array}$$

The corresponding codons (or vertices/nodes of a graph) of this matrix are {CCC, CCG, CGG, CGC, GGC, GGG, GCG, GCC}.

$$P_2 = \begin{array}{|c|c|c|c|c|c|c|c|} \hline 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ \hline 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ \hline 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ \hline \end{array}$$

The corresponding codons (or vertices/nodes of a graph) of P_2 *are* {CCA, CCU, CGA, CGU, GGA, GGU, GCU, GCA}.

$$P_3 = \begin{array}{|c|c|c|c|c|c|c|c|} \hline 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ \hline 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ \hline 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ \hline \end{array}$$

The corresponding codons (or vertices/nodes of a graph) of this matrix are {CAC, CAG, CUG, CUC, GUC, GUG, GAG, GAC}.

$$P_4 = \begin{array}{|c|c|c|c|c|c|c|c|} \hline 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ \hline 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ \hline 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \hline \end{array}$$

The corresponding codons (or vertices/nodes of a graph) of P_4 *are* {ACC, ACG, AGG, AGC, UGC, UGG, UCG, UCC}.

$$P_5 = \begin{array}{|c|c|c|c|c|c|c|c|}\hline 0&0&0&0&0&0&1&0\\\hline 0&0&0&0&0&0&0&1\\\hline 0&0&0&0&1&0&0&0\\\hline 0&0&0&0&0&1&0&0\\\hline 0&0&1&0&0&0&0&0\\\hline 0&0&0&1&0&0&0&0\\\hline 1&0&0&0&0&0&0&0\\\hline 0&1&0&0&0&0&0&0\\\hline \end{array}$$

The corresponding codons (or vertices/nodes of a graph) of this matrix are {ACA, ACU, AUG, AGA, UGA, UGU, UCU, UCA}.

$$P_6 = \begin{array}{|c|c|c|c|c|c|c|c|}\hline 0&0&1&0&0&0&0&0\\\hline 0&0&0&1&0&0&0&0\\\hline 1&0&0&0&0&0&0&0\\\hline 0&1&0&0&0&0&0&0\\\hline 0&0&0&0&0&0&1&0\\\hline 0&0&0&0&0&0&0&1\\\hline 0&0&0&0&1&0&0&0\\\hline 0&0&0&0&0&1&0&0\\\hline \end{array}$$

The corresponding codons (or vertices/nodes of a graph) of matrix P_6 *are* {CAA, CAU, CUU, CUA, GUA, GUU, GAU, GAA}.

$$P_7 = \begin{array}{|c|c|c|c|c|c|c|c|}\hline 0&0&0&0&1&0&0&0\\\hline 0&0&0&0&0&1&0&0\\\hline 0&0&0&0&0&0&1&0\\\hline 0&0&0&0&0&0&0&1\\\hline 1&0&0&0&0&0&0&0\\\hline 0&1&0&0&0&0&0&0\\\hline 0&0&1&0&0&0&0&0\\\hline 0&0&0&1&0&0&0&0\\\hline \end{array}$$

The corresponding codons (or vertices/nodes of a graph) of this matrix P_7 *are* {AAC, AAG, AGU, AUC, UCC, UGG, AUG, UAC}.

$$P_8 = \begin{array}{|c|c|c|c|c|c|c|c|}\hline 0&0&0&0&0&1&0&0\\\hline 0&0&0&0&1&0&0&0\\\hline 0&0&0&0&0&0&0&1\\\hline 0&0&0&0&0&0&1&0\\\hline 0&1&0&0&0&0&0&0\\\hline 1&0&0&0&0&0&0&0\\\hline 0&0&0&1&0&0&0&0\\\hline 0&0&1&0&0&0&0&0\\\hline \end{array}$$

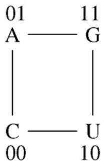

Figure 4. Hypercube when $n = 1$.

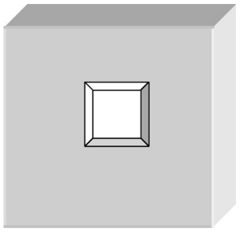

Figure 5. Hypercube when $n = 2$.

The corresponding codons (or vertices/nodes of a graph) of the 8*th matrix* P_8 *are* {AAA, AAU, AUU, AUA, UUA, UUU, UAU, UAA}.

Each permutation matrix is also doubly stochastic and symmetric.

Each matrix can be viewed as a vertex of a hypercube. We illustrate this hypercube for $n = 1, 2,$ and 3.

For the case $n = 1$, the hypercube is a square as shown in Fig. 4.

For $n = 2$, the hypercube consists of 16 vertices of {CC, CU, UU, UC, CA, CG, UG, UA, AA, AG, GG, GA, AC, AU, GU, GC} as illustrated in Fig. 5.

We include the hypercube in the case when $n = 3$ (Fig. 6) as illustrated in Petoukhov (2001).

As we have noted, the n-cube has a Hamiltonian cycle for $n = 2, 3, \ldots$. A Hamiltonian cycle is defined as a cycle in a graph G that contains each vertex in G exactly once, except for the starting and ending vertex that appears. It has been studied extensively in mathematics, physics and computer science. If we consider each entry of the genetic code-based matrix as a vertex of a graph G and (s_i, s_{i+1}), $i = 1, 2, \ldots, 2^n - 1$, then there is a Hamiltonian cycle in the graph G. This conclusion provides a pathway for the genetic code structure.

3.2. *Powers of the genetic matrix.* We next recall a well-known result on the power of the matrix. If A is the adjacency matrix of a simple graph, the ijth entry of A^m is equal to the number of paths of length m from vertex i to vertex j, $m = 1, 2, 3, \ldots$. To apply this result to the matrix D_{83}, we conclude that the number of paths of length m is equal to the entries of mth power of an adjacency matrix D_{83} corresponding to a simple graph with codons as vertices.

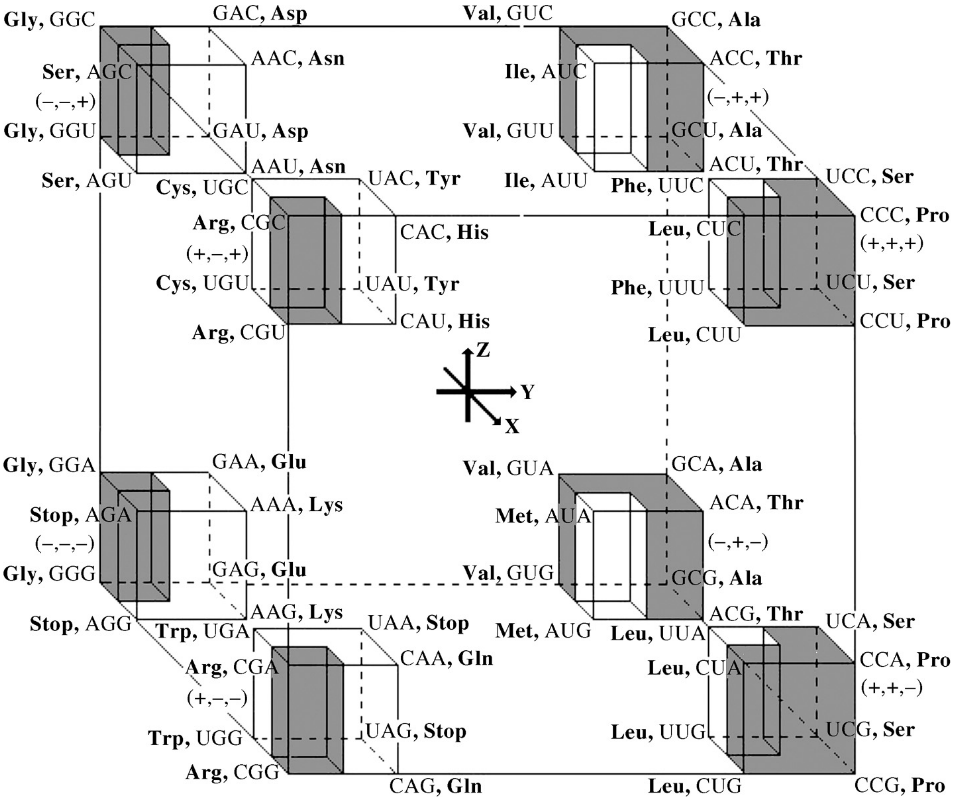

Figure 6. Hypercube when $n = 3$.

For $m = 1, 2, 3, \ldots$, we denote $D_{83}{}^{m}$ the mth power of matrix $G(i, j)$. It is easy to see that the matrices $D_{83}{}^{1}, D_{83}{}^{2}, \ldots, D_{83}{}^{m}$ are doubly stochastic, their eigenvalues are $\{(\lambda_1)^m, (\lambda_2)^m, \ldots, (\lambda_8)^m\} = \{0, 0, 0, 0, (-4)^m, (-4)^m, (-4)^m, 12^m\}$ with the same eigenvectors of D_{83}.

Here we illustrate the powers of matrix D_{83} when $m = 2$, and 3, respectively.

$(D_{83})^2 =$

24	20	16	20	16	12	16	20
20	24	20	16	12	16	20	16
16	20	24	20	16	20	16	12
20	16	20	24	20	16	12	16
16	12	16	20	24	20	16	20
12	16	20	16	20	24	20	16
16	20	16	12	16	20	24	20
20	16	12	16	20	16	20	24

The next iteration is the 3rd power of matrix D_{83}. The resulting matrix is

$$(D_{83})^2 = \begin{array}{|c|c|c|c|c|c|c|c|}
\hline
192 & 208 & 224 & 208 & 240 & 224 & 224 & 208 \\ \hline
208 & 192 & 208 & 224 & 224 & 240 & 208 & 224 \\ \hline
224 & 208 & 192 & 208 & 224 & 208 & 240 & 224 \\ \hline
208 & 224 & 208 & 192 & 208 & 224 & 224 & 240 \\ \hline
240 & 224 & 224 & 208 & 192 & 208 & 224 & 208 \\ \hline
224 & 240 & 208 & 224 & 208 & 192 & 208 & 224 \\ \hline
224 & 208 & 240 & 224 & 224 & 208 & 192 & 208 \\ \hline
208 & 224 & 224 & 240 & 208 & 224 & 208 & 192 \\ \hline
\end{array}$$

As the power m increases, the number of paths increases rapidly. However each iteration of the matrix generates only four different numbers due to the fact that the initial matrix $\mathbf{D_{83}}$ has Hamming distances of 0, 1, 2, and 3. One can extend this result into the general case of matrix $\mathbf{D_{2^n n}}$. If the length of DNA/RNA sequences is n, then all possible Hamming distances among the entries of the matrix $\mathbf{D_{2^n n}}$ are $0, 1, 2, \ldots, n$. The dimension of this matrix is 2^n by 2^n. Each entry of the matrix is a chain of DNA/RNA bases of length n. The iterations of the matrices provide a way of knowing the number of paths traveling from one entry to another within the matrix.

4. Conclusions

Our study showed a close relation between the genetic code and doubly stochastic matrix by using Hamming distance via the Gray code correspondence. The Hamming distance is applicable to encoded information, and is a particularly simple metric of comparison for error detections. The matrices are storages of digital data. The matrices appear in various dimensions with different shapes. Stochastic matrices motivated by language of probability show up repeatedly in nature. The biological evolution can be interpreted as a process of deployment and duplicating of certain forms of ordering. The digital information that underlies biochemistry, cell biology, and development can be represented by a simple strings of A's (adenine), C's (cytosine), G's (guanine) and T's (thymine) {or uracil (U) in RNA}. Linear sequences of these four letters on strings of molecules of heredity (DNA and RNA) contain the genetic information for protein synthesis in all living bodies from bacteria up to a whale and from a worm up to a bird and man. Formally we view these four letters as the elementary alphabet of a genetic code. This string is the root data structure of an organism's biology. Duplication with modification is the central paradigm of protein evolution, wherein new proteins and/or new biological functions are fashioned from earlier ones. Having advanced in the understanding of structurally functional features of the base systems of genetic coding, mankind extracts simultaneously an opportunity to advance in different areas of biology, which are built in the consent with these base systems. The structured set of triplets, allowing construction of the biperiodic table of genetic code, seems to be a generator of order on higher levels of the genetic system in many respects. It is hoped that relationships among genetic code, Hamming distance, and stochastic matrices will help us explore the structure of genetic code.

ACKNOWLEDGEMENTS

Funding for this project was provided through the Nova Southeastern University President's Faculty Research and Development Grant No. 335358.

REFERENCES

He, M. (2003a). Genetic code, attributive mappings and stochastic matrices. *Bull. Math. Biol.* (in press) (doi: 10.1016/j.bulm.2003.10.002).

He, M. (2003b). Double helical sequences and doubly stochastic matrices. *Symmetry: Culture and Science: Symmetries in Genetic Information* (in press).

He, M. (2003c). 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.

Jimenéz-Montaño, M. A., C. R. Mora-Basáñez and T. Pöschel (1994). On the hypercube structure of the genetic code, in *Proc. 3. Int. Conf. on Bioinformatics and Genome Research*, A. H. Lim and A. C. Cantor (Eds), World Scientific, p. 445.

Knight, R. D., S. J. Freeland and L. F. Landweber (1999). Selection, history and chemistry: the three faces of the genetic code. *TIBS* **24**, 241–247.

Petoukhov, S. V. (1999). Genetic code and the ancient Chinese book of changes. *Symmetry: Culture Sci.* **10**, 211–226.

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].

Štambuk, N. (2000). Universal metric properties of the genetic code. *Croatica Chemica ACTA* **73**, 1123–1139.

Swanson, R. (1984). A unifying concept for the amino acid code. *Bull. Math. Biol.* **46**, 187–203.

Yang, C. M. (2003). The naturally designed spherical symmetry in the genetic code (manuscript).

Received 4 November 2003 and accepted 23 January 2004