Studying Genetic Code by a Matrix Approach

Tanner Crowder [1] and Chi-Kwong Li[2]

Department of Mathematics, The College of William and Mary,

Williamsburg, Virginia 23185, USA

E-mails: tjcrow@wm.edu, ckli@math.wm.edu.

Abstract

Following Petoukhov and his collaborators we use two length n zero-one sequences, α and β, to represent a length n genetic sequence $\binom{\alpha}{\beta}$ so that the columns of $\binom{\alpha}{\beta}$ have the following correspondence with the nucleotides: $C \sim \binom{0}{0}$, $U \sim \binom{1}{0}$, $G \sim \binom{1}{1}$, $A \sim \binom{0}{1}$. Using the Gray code ordering to arrange α and β, we build a $2^n \times 2^n$ matrix C_n including all the 4^n length n genetic sequences. Furthermore, we use the Hamming distance of α and β to construct a $2^n \times 2^n$ matrix D_n. We explore structures of these matrices, refine the results in earlier papers, and propose new directions for further research.

1 Introduction

Genetic Code is the set of rules by which information encoded in RNA/DNA is translated into amino acid sequences in living cells. The bases for the encoded information are nucleotides. There are four nucleotide bases for RNA: Adenine, Uracil, Guanine, and Cytosine, which are labeled by A, U, G, and C respectively, (in DNA Uracil is replaced by Thymine (T)). In canonical genetic code, codons are tri-nucleotide sequences such that each triplet relates to an amino acid. For example, the codon CAG encodes the amino acid Glutamine. Amino acids are the basic building blocks of proteins.

The genetic code was cracked by Holley, Khorana, Nirenberg and co-workers in the sixties. It stimulated interest of other researchers to study how genetic code was translated into amino acids. There are 20 different amino acids (plus start and stop codons), and since there are four nucleotide bases, A, U, C, and G, there are 4^n different combinations of bases, for a string of length n. Therefore, $n = 3$ is the smallest number of bases that could be used to represent the 20 different amino acids. There is degeneracy between the codons, i.e., more than one codon can represent the same amino acid; however, two different amino acids cannot be represented by the same codon.

In general, genetic sequences are very long, so it is difficult to extract information or to observe patterns. The focus of this study is examining matrices which will contain all length n nucleotide sequences and building matrices that can efficiently represent the genetic sequences. Many studies have been devoted to examining how genetic code has evolved. Patterns that arise in genetic code suggest that genetic code evolved to minimize the effects of mutations; for example, see [1, 6, 5, 16]. One current aspect of research is examining the redundancy of genetic code and it's effect on the dynamic of evolution [3]. In connection to this we consider a graph $G = (V, E)$, where V is the set of all length n genetic sequences, and E is the edge set where two vertices are adjacent if they differ

[1]Research of this author was partially supported by the NSF CSUMS and NSF UBM undergraduate research grants at William and Mary; this research was done while he was a student at William and Mary. His current address is: US Navel Research Laboratory, 4555 Overlook Ave. S.W., Washington, DC 20375.

[2]Li is the corresponding author. He is an honorary professor of the University of Hong Kong. His research was partially supported by NSF and the William and Mary Plumeri Award.

by one nucleotide base. A *Hamilton circuit* will be given for that graph which may help analyze mutations in genetic code.

Swanson [18] suggested that each nucleotide could be represented as a Gray code sequence. Gray code is an encoding scheme with the property that two consecutive sequences only differ by one position [19]. For example, the classical binary representations for three and four are 011 and 100 respectively, but the Gray code representations for three and four are 011 and 010, respectively. In classical binary, 011 and 100 differ in all three positions, but in the Gray code representation 011 and 010 differ in only one position, namely the last position.

Define G_n to be all the Gray code sequences of length n, which can be generated by a recursive algorithm. G_n is constructed by taking the sequences from G_{n-1} and prepending a 0 to them then taking the sequences of G_{n-1} in reverse order and prepending a 1 to them; therefore $G_n = \{0||a_0, 0||a_1, \ldots, 0||a_{n-1}, 1||a_{n-1}, 1||a_{n-2}, \ldots, 1||a_0\}$, where $a_i \in G_{n-1}$. Note $a||b$ is the operation a concatenate b. To illustrate this process take $G_1 = \{0, 1\}$. Then by construction $G_2 = \{0||0, 0||1, 1||1, 1||0\} = \{00, 01, 11, 10\}$.

Initially Gray code was intended for transmitting information where a change in one bit would distort the information less than if the information was encoded using the standard binary representation [19]. It is natural to represent genetic code in this manner because Gray code is designed to minimize the mismatches between the digit encoding adjacent bases and therefore minimizing the mismatches between nearby chromosome segments. This may help study the mutation occurring in genetic sequences [7, 8].

Following He et al. [8], we use the following correspondence for the nucleotides and two-bit Gray codes: $C \sim \binom{0}{0}$, $U \sim \binom{1}{0}$, $G \sim \binom{1}{1}$, and $A \sim \binom{0}{1}$. The genetic code-based matrix, which will contain all nucleotide strings of length n is defined as C_n. The Gray code sequences represented by C_n will be denoted by a $2^n \times 2^n$ matrix. Here are C_1, C_2, C_3 and their corresponding Gray code representations.

$$C_1 = \begin{pmatrix} C & U \\ A & G \end{pmatrix} \sim \begin{matrix} \\ 0 \\ 1 \end{matrix} \begin{matrix} \begin{matrix} 0 & 1 \end{matrix} \\ \begin{pmatrix} \binom{0}{0} & \binom{1}{0} \\ \binom{0}{1} & \binom{1}{1} \end{pmatrix} \end{matrix};$$

$$C_2 = \begin{pmatrix} CC & CU & UU & UC \\ CA & CG & UG & UA \\ AA & AG & GG & GA \\ AC & AU & GU & GC \end{pmatrix} \sim \begin{matrix} \\ 00 \\ 01 \\ 11 \\ 10 \end{matrix} \begin{matrix} \begin{matrix} 00 & 01 & 11 & 10 \end{matrix} \\ \begin{pmatrix} \binom{00}{00} & \binom{01}{00} & \binom{11}{00} & \binom{10}{00} \\ \binom{00}{01} & \binom{01}{01} & \binom{11}{01} & \binom{10}{01} \\ \binom{00}{11} & \binom{01}{11} & \binom{11}{11} & \binom{10}{11} \\ \binom{00}{10} & \binom{01}{10} & \binom{11}{10} & \binom{10}{10} \end{pmatrix} \end{matrix};$$

$$C_3 = \begin{pmatrix} CCC & CCU & CUU & CUC & UUC & UUU & UCU & UCC \\ CCA & CCG & CUG & CUA & UUA & UUG & UCG & UCA \\ CAA & CAG & CGG & CGA & UGA & UGG & UAG & UAA \\ CAC & CAU & CGU & CGC & UGC & UGU & UAU & UAC \\ AAC & AAU & AGU & AGC & GGC & GGU & GAU & GAC \\ AAA & AAG & AGG & AGA & GGA & GGG & GAG & GAA \\ ACA & ACG & AUG & AUA & GUA & GUG & GCG & GCA \\ ACC & ACU & AUU & AUC & GUC & GUU & GCU & GCC \end{pmatrix}$$

$$\sim \begin{array}{c} \\ 000 \\ 001 \\ 011 \\ 010 \\ 110 \\ 111 \\ 101 \\ 100 \end{array} \begin{array}{c} \begin{array}{cccccccc} 000 & 001 & 011 & 010 & 110 & 111 & 101 & 100 \end{array} \\ \begin{pmatrix} \binom{000}{000} & \binom{001}{000} & \binom{011}{000} & \binom{010}{000} & \binom{110}{000} & \binom{111}{000} & \binom{101}{000} & \binom{100}{000} \\ \binom{000}{001} & \binom{001}{001} & \binom{011}{001} & \binom{010}{001} & \binom{110}{001} & \binom{111}{001} & \binom{101}{001} & \binom{100}{001} \\ \binom{000}{011} & \binom{001}{011} & \binom{011}{011} & \binom{010}{011} & \binom{110}{011} & \binom{111}{011} & \binom{101}{011} & \binom{100}{011} \\ \binom{000}{010} & \binom{001}{010} & \binom{011}{010} & \binom{010}{010} & \binom{110}{010} & \binom{111}{010} & \binom{101}{010} & \binom{100}{010} \\ \binom{000}{110} & \binom{001}{110} & \binom{011}{110} & \binom{010}{110} & \binom{110}{110} & \binom{111}{110} & \binom{101}{110} & \binom{100}{110} \\ \binom{000}{111} & \binom{001}{111} & \binom{011}{111} & \binom{010}{111} & \binom{110}{111} & \binom{111}{111} & \binom{101}{111} & \binom{100}{111} \\ \binom{000}{101} & \binom{001}{101} & \binom{011}{101} & \binom{010}{101} & \binom{110}{101} & \binom{111}{101} & \binom{101}{101} & \binom{100}{101} \\ \binom{000}{100} & \binom{001}{100} & \binom{011}{100} & \binom{010}{100} & \binom{110}{100} & \binom{111}{100} & \binom{101}{100} & \binom{100}{100} \end{pmatrix} \end{array}.$$

When $n = 3$, or is a multiple of 3, C_n contains nucleotide triplets, which are codons. Therefore interesting biological structure starts to appear in C_3.

The Hamming distance is a measure of how many positions are different in two equal length sequences. For example, the binary sequences 001 and 011 have a Hamming distance 1, since there is only one difference in the second position. This is precisely the Hamming distance of the two binary sequences corresponding to the codon CAG because $CAG \sim \binom{001}{011}$–by construction. The Hamming distance is not exclusive to binary sequences; the words "math" and "bath" have a Hamming distance 1 because they differ in the first position. To get a better understanding of the Genetic code matrix and the recursion, the Hamming distance matrices, D_n, associated with C_n will be studied. Each entry of D_n is the Hamming distance between the Gray code sequences that represent the nucleotides of C_n. For example, D_1, D_2, D_3 are as follows:

$$D_1 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad D_2 = \begin{pmatrix} 0 & 1 & 2 & 1 \\ 1 & 0 & 1 & 2 \\ 2 & 1 & 0 & 1 \\ 1 & 2 & 1 & 0 \end{pmatrix}, \quad D_3 = \begin{pmatrix} 0 & 1 & 2 & 1 & 2 & 3 & 2 & 1 \\ 1 & 0 & 1 & 2 & 3 & 2 & 1 & 2 \\ 2 & 1 & 0 & 1 & 2 & 1 & 2 & 3 \\ 1 & 2 & 1 & 0 & 1 & 2 & 3 & 2 \\ 2 & 3 & 2 & 1 & 0 & 1 & 2 & 1 \\ 3 & 2 & 1 & 2 & 1 & 0 & 1 & 2 \\ 2 & 1 & 2 & 3 & 2 & 1 & 0 & 1 \\ 1 & 2 & 3 & 2 & 1 & 2 & 1 & 0 \end{pmatrix}.$$

The Hamming distance matrix gives information about genetic code and yet requires less storage. Specifically it gives information about the composition of each entry in C_n. It shows how many possible U or A and C or G nucleotides are contained in each entry. However, it only shows how many total U's and A's (and therefore C's and G's) appear combined.

In the following discussion, we always assume that C_n and D_n are defined as in this section. Also, we let F_n be the $2^n \times 2^n$ matrix with (i, j) entry equal to 1 if $i + j = 2^n + 1$, and all other entries equal to 0 (this will also be referred to the anti-diagonal matrix); we let J_n be the $2^n \times 2^n$ matrix with all entries equal to 1. For example, we have

$$F_1 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad F_2 = \begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix}; \quad J_1 = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}, \quad J_2 = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{pmatrix}.$$

Also we write $X \oplus Y = \begin{pmatrix} X & 0 \\ 0 & Y \end{pmatrix}$, for two square matrices.

Although our study does not have direct applications to biological problems yet, it is our hope that the matrix results will help organize, store, and retrieve data in a systematic way so that hidden patterns in genetic sequences can be recognized effectively by computers or humans. As pointed out by a referee, information in computers is often stored in matrix form so that matrix methods are used intensively to provide noise-immunity of information transfer of digital data. In view of this, matrix techniques have been used in molecular genetics and in bioinformatics in the last decade. Also, it has been suggested in [15] that bimolecular computation can be utilized by applying biotechnology operation to do computation. In this setting, data would be encoded using a basis of RNA or DNA. The Gray code representation of genetic sequences could be used as a potential mathematical basis/representation of RNA/DNA computation. Furthermore, it is worth mentioning that graph theoretic approach has also been used to understand different RNA/DNA sequences and gene families; see for example [4, 12, 13].

Our paper is organized as follows. We obtain some basic properties for the matrices C_n and D_n in Section 2. Section 3 concerns the eigenstructure of the matrix D_n. In particular, we show that D_n admits a spectral decomposition $n2^{n-1}v_0v_0^* - 2^{n-1}\sum_{j=1}^{n} v_jv_j^*$, where $\{v_0, \ldots, v_n\}$ is an orthonormal set in $\mathbf{R}^N$ with $N = 2^n$. Using this result, one can evaluate the powers of the matrix D_n. In Section 4, we obtain decomposition of D_n related to some graph structure of genetic sequences. Future research directions and additional remarks are given in Section 5.

2 Properties of C_n and D_n

Let C_n and D_n be defined as in Section 1. We first present an easy recurrence construction of the matrices. Moreover, we show that D_n is *bisymmetric*, i.e., $F_nD_nF_n = D_n^t = D_n$.

In the following theorem, for $Z \in \{C, U, A, G\}$, $Z||C_n$ denotes the $2^n \times 2^n$ matrix obtained by prepending Z to each sequence in C_n; C_nF_n denotes the matrix obtained from C_n by arranging its columns in the reverse order; F_nC_n denotes the matrix obtained from C_n by arranging its rows in the reverse order, etc; in other words, F_n is a permutation matrix.

Theorem 2.1 *Suppose C_n and D_n are defined as in Section 1. Then*

$$C_{n+1} = \begin{pmatrix} C||C_n & U||C_nF_n \\ A||F_nC_n & G||F_nC_nF_n \end{pmatrix}.$$

If

$$D_n = \begin{pmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{pmatrix},$$

where B_{ij} is a $2^{n-1} \times 2^{n-1}$ submatrix, then $B_{11} = B_{22} = D_{n-1}$, $B_{12} = B_{21}$, and both B_{ij} and D_n are bisymmetric; moreover,

$$D_{n+1} = \begin{pmatrix} B_{11} & B_{12} & 2J_{n-1} + B_{11} & B_{12} \\ B_{12} & B_{11} & B_{12} & 2J_{n-1} + B_{11} \\ 2J_{n-1} + B_{11} & B_{12} & B_{11} & B_{12} \\ B_{12} & 2J_{n-1} + B_{11} & B_{12} & B_{11} \end{pmatrix}.$$

Proof. Clearly, we have

$$C_1 = \begin{pmatrix} C & U \\ A & G \end{pmatrix} \quad \text{and} \quad C_2 = \begin{pmatrix} CC & CU & UU & UC \\ CA & CG & UG & UA \\ AA & AG & GG & GA \\ AC & AU & GU & GC \end{pmatrix}.$$

In general, for $n \geq 2$, suppose we have constructed C_n such that each entry of C_n is encoded by $\alpha = (a_1 \cdots a_n), \beta = (b_1 \cdots b_n) \in G_n$, where G_n denotes the set of Gray code sequences of length n. We refer to it as the (β, α) entry in C_n.

Suppose

$$C_{n+1} = \begin{pmatrix} X_1 & X_2 \\ X_3 & X_4 \end{pmatrix}.$$

Then the entries of X_1 are encoded by two Gray code sequences in G_{n+1} identified as $\begin{pmatrix} \hat{\alpha} \\ \hat{\beta} \end{pmatrix}$ with $\hat{\alpha} = 0||\alpha$ and $\hat{\beta} = 0||\beta$, where $\alpha, \beta \in G_n$. Hence, X_1 is obtained from C_n by prepending $C \sim \begin{pmatrix} 0 \\ 0 \end{pmatrix}$ to the beginning of the genetic sequences.

Similarly, if $\alpha_1, \ldots, \alpha_{2^n}$ are the Gray code sequences of length n, and $\tilde{\alpha}_j = 1||\alpha_j \in G_{n+1}$ for $j = 1, \ldots, 2^n$. Then the entries of X_2 in the row labeled by the Gray code sequence $\tilde{\beta} = 0||\beta$, with $\beta \in G_n$, have the form

$$\begin{pmatrix} \tilde{\alpha}_{2^n} \\ \tilde{\beta} \end{pmatrix}, \begin{pmatrix} \tilde{\alpha}_{2^n-1} \\ \tilde{\beta} \end{pmatrix}, \ldots, \begin{pmatrix} \tilde{\alpha}_1 \\ \tilde{\beta} \end{pmatrix}.$$

Consequently, if we arrange the columns of C_n in the reverse order to get the matrix C_nF_n, and added $U \sim \begin{pmatrix} 1 \\ 0 \end{pmatrix}$ to the left most position of each entry of C_nF_n, we obtain the submatrix X_2.

We can use a similar argument to conclude that $X_3 = A||F_nC_n$ and $X_4 = G||F_nC_nF_n$ as asserted.

Now, consider D_n. The bisymmetric structure is clear for D_1, D_2, and one can construct D_2 from D_1 as asserted. We will show that the D_{n+1} can be constructed from D_n and D_{n-1} as asserted for $n \geq 2$ by induction. The bisymmetry condition on D_{n+1} will follow from the construction.

Assume that the assertion is true for $D_1, D_2, \ldots, D_n$ with $n \geq 2$. Consider

$$D_{n+1} = \begin{pmatrix} X_{11} & X_{12} & X_{13} & X_{14} \\ X_{21} & X_{22} & X_{23} & X_{24} \\ X_{31} & X_{32} & X_{33} & X_{34} \\ X_{41} & X_{42} & X_{43} & X_{44} \end{pmatrix} \quad \text{where } X_{ij} \in M_{2^{n-1}}.$$

Note that each entry of D_{n+1} corresponds to a sequence in C_{n+1} represented as $\begin{pmatrix} \tilde{\alpha} \\ \tilde{\beta} \end{pmatrix}$, where $\tilde{\alpha}, \tilde{\beta} \in G_{n+1}$, and the corresponding entry in D_{n+1} is just the Hamming distance between $\tilde{\alpha}$ and $\tilde{\beta}$. Now, each entry in

$$\begin{pmatrix} X_{11} & X_{12} \\ X_{21} & X_{22} \end{pmatrix}$$

corresponds to an entry in $C||C_n \sim \begin{pmatrix} 0||\alpha \\ 0||\beta \end{pmatrix}$ by the result on C_{n+1} with $\alpha, \beta \in G_n$. Clearly, the Hamming distance between $0||\alpha, 0||\beta \in G_{n+1}$ is the same as that between $\alpha, \beta \in G_n$. It follows

that

$$D_n = \begin{pmatrix} X_{11} & X_{12} \\ X_{21} & X_{22} \end{pmatrix}.$$

Similarly, we can use the result on C_{n+1} to show that

$$F_n D_n F_n = \begin{pmatrix} X_{33} & X_{34} \\ X_{43} & X_{44} \end{pmatrix}.$$

By induction assumption, D_n is bisymmetric. We conclude that $F_n D_n F_n = D_n$.

Next, we consider

$$\begin{pmatrix} X_{13} & X_{14} \\ X_{23} & X_{24} \end{pmatrix}.$$

Note that if $(a_1, \ldots, a_{n-2}), (b_1, \ldots, b_{n-2}) \in G_{n-2}$, we can label the entries of D_{n+1} as follows.

$$D_{n+1} = \begin{array}{c} \\ (00b_1 \cdots b_{n-2}) \\ (01b_{n-2} \cdots b_1) \\ (11b_1 \cdots b_{n-2}) \\ (10b_{n-2} \cdots b_1) \end{array} \begin{array}{c} \begin{array}{cccc} (00a_1 \cdots a_{n-2}) & (01a_{n-2} \cdots a_1) & (11a_1 \cdots a_{n-2}) & (10a_{n-2} \cdots a_1) \end{array} \\ \begin{pmatrix} X_{11} & X_{12} & X_{13} & X_{14} \\ X_{21} & X_{22} & X_{23} & X_{24} \\ X_{31} & X_{32} & X_{33} & X_{34} \\ X_{41} & X_{42} & X_{43} & X_{44} \end{pmatrix} \end{array}.$$

Compare two entries in X_{11} and X_{13} lying in the same row and column labeled by $\alpha', \beta' \in G_{n-1}$. Then these entries are labeled by $\begin{pmatrix} 00||\alpha' \\ 00||\beta' \end{pmatrix}$ and $\begin{pmatrix} 11||\alpha' \\ 00||\beta' \end{pmatrix}$, respectively, in C_{n+1}. These two entries from C_{n+1} have Hamming distances differing by 2. So, we see that $X_{13} = X_{11} + 2J_{n-1}$.

Similarly, consider the two $2^{n-1} \times 2^{n-1}$ matrices X_{12} and X_{14} and compare their entries in X_{12} and X_{14} lying in the row and column labeled by $\alpha', \beta' \in G_{n-1}$. Then these entries are labeled by $\begin{pmatrix} 10||\alpha' \\ 00||\beta' \end{pmatrix}$ and $\begin{pmatrix} 01||\alpha' \\ 00||\beta' \end{pmatrix}$ respectively, in C_{n+1}. These two entries from C_{n+1} have the same Hamming distance. So, we see that $X_{12} = X_{14}$.

We can apply similar arguments to show that $X_{21} = X_{23}$, $X_{24} = X_{22} + 2J_{n-1}$, $X_{31} = X_{33} + 2J_{n-1}$, $X_{32} = X_{34}$, $X_{41} = X_{43}$ and $X_{42} = X_{44} + 2J_{n-1}$. By induction assumption, D_n is bisymmetric, one sees that each X_{ij} is bisymmetric and so is D_{n+1}. Our result follows. □

Using Theorem 2.1, we can refine [8, Theorem A], which was stated without proof. We begin with the following corollary covering [8, Theorem A (i)].

Corollary 2.2 *Let C_n and D_n be defined as in Section 1. In C_n two neighboring entries of genetic code in both directions differ by exactly one base; each two neighboring entries of D_n differ by one. Here the first entry and the last entry of a row (respectively, a column) in C_n or D_n is also considered as neighbors.*

Proof. By the Gray code construction each binary sequence in G_n differs by one from a neighboring binary sequence. Fix a row in C_n. The entries will be represented by $\binom{\alpha}{\beta}$ where $\alpha, \beta \in G_n$. If the row is fixed, β will stay constant for all the columns, and α will vary only by one position when moving from one column to another or from the last column to the first column. Thus the genetic sequence will change by one nucleotide when moving along a row. The same is true if the column is fixed. The conclusion on C_n follows.

Similarly the result can be proven for D_n as well. □

In [8, Theorem A (v)], the authors wrote $D_n = (B_{ij})_{1\le i,j\le 2^{n-1}}$ such that B_{ij} is 2×2 for each (i,j). They showed that there are 2×2 matrices $T_0,\dots,T_{n-1}$ such that $B_{ij}\in\{T_0,\dots,T_{n-1}\}$ for each (i,j), where $T_0 = D_1$ and T_j can be easily constructed from T_0. Moreover, they determined the frequency distribution of the matrices $T_0,\dots,T_{n-1}$ as entry of the block matrix $D_n = (B_{ij})_{1\le i,j\le 2^{n-1}}$. We have the following generalization.

Theorem 2.3 *Let D_n be defined as in Section 1. Suppose $D_n = (B_{ij})_{1\le i,j\le 2^m}$, where $B_{ij}\in M_{2^k}$ such that $m = n-k\ge 1$. Then there are $m+1$ distinct matrices $T_0, T_1,\dots,T_m$ in each row and each column of the block matrix $(B_{ij})_{1\le i,j\le 2^m}$ defined and arranged in D_n according to the following scheme:*

$T_0 = D_k$, $T_1 = T_0 + 2(J_{k-1}\oplus J_{k-1})$, and $T_{j+2} = T_j + 2J_k$ for $0\le j\le m-2$.

For $1\le i,j\le 2^m$, $B_{ij} = T_\ell$ if the (i,j) entry of D_m equals ℓ.

Consequently, in each row and each column, the matrix T_j will appear $\binom{m}{j}$ times.

Note that by Theorem 2.1, we can build D_m from D_1 in $m-1$ steps, and use some two step recurrence relations to define the entries of D_m. This theorem and its proof show that we can extend the procedures to build D_n from D_{k+1} in $m-1$ steps for $n = m+k$, and determine the $2^k\times 2^k$ submatrices B_{ij} of D_n by the same two step recurrence relations.

Proof. We prove the theorem by induction on m. Suppose $m = 1$. Then

$$D_n = \begin{pmatrix} B_{11} & B_{12}\\ B_{21} & B_{22}\end{pmatrix}$$

with $T_0 = D_{n-1}$ and $T_1 = T_0 + 2(J_{n-1}\oplus J_{n-1})$ as asserted.

Assume that $D_n = (B_{ij})_{1\le i,j\le 2^m}$ so that $B_{ij} = T_\ell$ is $2^k\times 2^k$ as asserted. By Theorem 2.1,

$$D_{n+1} = (\widetilde{B}_{ij})_{1\le i,j\le 2^{m+1}} = \begin{pmatrix} D_n & D_n + 2(J_{n-1}\oplus J_{n-1})\\ D_n + 2(J_{n-1}\oplus J_{n-1}) & D_n\end{pmatrix}.$$

First consider those $\widetilde{B}_{ij}$ with $1\le i,j\le 2^m$ and $2^m < i,j\le 2^{m+1}$. Evidently, for $r,s\in\{1,\dots,2^m\}$,

$$\widetilde{B}_{r+2^m,s+2^m} = \widetilde{B}_{rs} = B_{rs} = T_\ell,$$

where the index ℓ is the (r,s) entry of D_m. By the construction of D_{m+1} from D_m, we see that ℓ is the (r,s) entry and also the $(r+2^m, s+2^m)$ entry of D_{m+1}.

Next, we consider $\widetilde{B}_{ij}$ with

$$i\in\{1,\dots,2^m\} = \{r : 1\le r\le 2^{m-1}\}\cup\{r+2^{m-1} : 1\le r\le 2^{m-1}\},\qquad \text{and}$$

$$j\in\{2^m+1,\dots,2^{m+1}\} = \{s+2^m : 1\le s\le 2^{m-1}\}\cup\{s+2^{m-1}3 : 1\le s\le 2^{m-1}\}.$$

In other words, the submatrices of $D_n + 2(J_{n-1}\oplus J_{n-2})$ lying at the north east corner of D_{n+1}. Suppose $r,s\in\{1,\dots,2^{m-1}\}$. We have

$$\widetilde{B}_{r,s+2^m} = B_{rs} + 2J_k = T_\ell + 2J_k = T_{\ell+2},$$

where the index ℓ is the (r,s) entry of D_m. By the construction of D_{m+1} from D_m, we see that $\ell+2$ is the $(r,s+2^m)$ entry of D_{m+1}. Similarly, we see that

$$\widetilde{B}_{r+2^{m-1},s+2^{m-1}3} = B_{r+2^{m-1},s+2^{m-1}} + 2J_k = T_\ell + 2J_k = T_{\ell+2},$$

and $\ell+2$ is the $(r+2^{m-1}, s+2^{m-1}3)$ entry of D_{m+1}. Furthermore,

$$\widetilde{B}_{r,s+2^{m-1}3} = B_{r,s+2^{m-1}} = T_\ell,$$

where ℓ is the $(r, s+2^{m-1}3)$ entry of D_{m+1}. We also have

$$\widetilde{B}_{r+2^{m-1},s+2^m} = B_{r+2^{m-1},s} = T_\ell,$$

where ℓ is the $(r+2^{m-1}, s+2^m)$ entry of D_{m+1}. We can analyze the south west block of D_{m+1}, and conclude that $\widetilde{B}_{rs} = T_\ell$ whenever the (r,s) entry of D_{m+1} is ℓ. Hence, we obtain the statement concerning the arrangement of $T_0,\ldots,T_m$ in each row and each column of D_n.

Finally, note that the number of T_ℓ appearing in each row (or column) of $(B_{ij})_{1\le i,j\le 2^m}$ is the same as the number of ℓ in each row (or column) of D_m. Each entry of D_m is Hamming distance of $\alpha,\beta\in G_m$, where $\binom{\alpha}{\beta}$ corresponds to the genetic sequence in C_m. For a fixed row of C_m, the genetic sequences are encoded by

$$\binom{\alpha_1}{\beta},\ldots,\binom{\alpha_{2^m}}{\beta},$$

where $\beta\in G_m=\{\alpha_1,\ldots,\alpha_{2^m}\}$. Clearly, the number of sequences in G_m differ with β in ℓ positions equals $\binom{m}{\ell}$ for $\ell=0,\ldots,m$. Hence, ℓ will occur $\binom{m}{\ell}$ times in each row of D_m for $\ell=0,\ldots,m$. So, T_j will occur $\binom{m}{\ell}$ times in each row of $(B_{ij})_{1\le i,j\le 2^m}$. The proof for the column is similar. □

In [8, Theorem A (v)], the authors showed that D_1 and D_2 are principal submatrices of D_n. By Theorem 2.3, we have the following extension.

Corollary 2.4 *Let $D_n=(B_{ij})_{1\le i,j\le 2^m}$, where $B_{ij}\in M_{2^k}$ such that $m=n-k\ge 1$. Then*

$$D_k = B_{11} = B_{22} = \cdots = B_{2^m,2^m}$$

and

$$D_{k+1} = \begin{pmatrix} B_{jj} & B_{j,j+1} \\ B_{j+1,j} & B_{j+1,j+1}\end{pmatrix},\quad j=1,\ldots,2^m-1.$$

Note that if $m=2$, i.e., D_n is a 4×4 block matrix, then we see that D_{n-1} is centrally embedded in D_n.

Putting $(m,k)=(n,0)$ in Theorem 2.3 and using Theorem 2.1, we have the following corollary; see [8, Theorem A (ii)–(iv)].

Corollary 2.5 *Let D_n be defined as in Section 1. Each row and each column of D_n has $\binom{n}{k}$ entries equal to k so that the row sum (respectively, column sum) equals $n2^{n-1}$. Consequently, the total sum of the entries of the matrix D_n is $n2^{2n-1}$, and $D_n/(2^{n-1}n)$ is a bisymmetric doubly stochastic matrices.*

3 Eigenstructure and powers of D_n

Theorem 3.1 *The matrix $D_n \in M_{2^n}$ has $n+1$ nonzero eigenvalues equal to*

$$n2^{n-1}, \overbrace{-2^{n-1}, -2^{n-1}, \ldots, -2^{n-1}}^{n}.$$

Proof. We will prove the theorem by induction. The result for $n = 1$ is clear. Assume the result is true for D_n. Clearly D_n has two unit eigenvectors of the form

$$x = 2^{-n/2}(1, 1, ..., 1)^t \quad \text{and} \quad y = 2^{-n/2}(\underbrace{1, \ldots, 1}_{2^{n-1}}, \underbrace{-1, \ldots, -1}_{2^{n-1}})^t$$

for the eigenvalues $n2^{n-1}$ and -2^{n-1}. By induction assumption, there is an orthogonal matrix P, with x and y as the first two columns such that

$$A_n = P^t D_n P = [n2^{n-1}] \oplus (-2^{n-1}) I_n \oplus 0_{2^n - n - 1}.$$

Let $Q = P \oplus P$. Then

$$\begin{aligned} Q^t D_{n+1} Q &= Q^t \begin{pmatrix} D_n & D_n \\ D_n & D_n \end{pmatrix} Q + Q^t \begin{pmatrix} 0 & 0 & 2J_{n-1} & 0 \\ 0 & 0 & 0 & 2J_{n-1} \\ 2J_{n-1} & 0 & 0 & 0 \\ 0 & 2J_{n-1} & 0 & 0 \end{pmatrix} Q \\ &= \begin{pmatrix} A_n & A_n \\ A_n & A_n \end{pmatrix} + \begin{pmatrix} 0 & C_n \\ C_n & 0 \end{pmatrix}, \end{aligned}$$

where $C_n = \operatorname{diag}(2^n, 2^n, 0, \ldots, 0)$.

Up to a permutation similarity, $Q^t D_{n+1} Q$ is a direct sum: $R_1 \oplus R_2 \oplus R_3 \oplus 0_{2^{n+1} - 2n - 2}$, where

$$R_1 = 2^{n-1} \begin{pmatrix} n & n+2 \\ n+2 & n \end{pmatrix}, \quad R_2 = 2^{n-1} \begin{pmatrix} -1 & 1 \\ 1 & -1 \end{pmatrix}$$

and R_3 is a direct sum of $(n-1)$ copies of the matrix

$$-2^{n-1} \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}.$$

Notice $R_1 \oplus R_2$ has eigenvalues $(n+1)2^n, -2^n, -2^n, 0$, and all the $n-1$ nonzero eigenvalues of R_3 are equal to -2^n. By an inductive argument, the assertion follows. □

Next we obtain an orthonormal set of eigenvectors of D_n which correspond to the nonzero eigenvalues.

Theorem 3.2 *An orthonormal set of eigenvectors of D_n corresponding to the nonzero eigenvalues $n2^{n-1}, -2^{n-1}, \ldots, -2^{n-1}$ can be constructed as follows. For D_1, the orthonormal eigenvectors are $\frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ 1 \end{pmatrix}$ and $\frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ -1 \end{pmatrix}$. Suppose $v_0, v_1, \ldots, v_n$ is constructed for D_n. Then*

$$\tilde{v}_j = \frac{1}{\sqrt{2}} \begin{pmatrix} v_j \\ v_j \end{pmatrix} \text{ for } j = 0, \ldots, n \quad \text{and } \tilde{v}_{n+1} = \frac{1}{\sqrt{2}} \begin{pmatrix} v_0 \\ -v_0 \end{pmatrix},$$

form an orthonormal set of eigenvectors of D_{n+1} corresponding to the nonzero eigenvalues.

Proof. The results can be verified for $n = 1, 2$. Suppose $n > 2$, and the result is true for D_m with $m \leq n$. By Corollary 2.5, $(n+1)2^n$ is the common row sum of D_{n+1}. Thus, $D_{n+1}\tilde{v}_0 = (n+1)2^n\tilde{v}_0$.

Let $K_n = J_{n-1} \oplus J_{n-1} \in M_{2^n}$. By induction assumption, $v_0, \ldots, v_n$ form an orthonormal set of eigenvectors for D_n. It can be seen that $K_n v_j = 0$ for all $j = 1, \ldots, n$. Thus,

$$D_{n+1}\tilde{v}_j = \frac{1}{\sqrt{2}}\begin{pmatrix} 2D_n v_j \\ 2D_n v_j \end{pmatrix} = \frac{1}{\sqrt{2}}\begin{pmatrix} 2 \cdot -2^{n-1}v_j \\ 2 \cdot -2^{n-1}v_j \end{pmatrix} = -2^n\tilde{v}_j \qquad j = 1, \ldots, n.$$

Moreover,

$$D_{n+1}\tilde{v}_{n+1} = \frac{1}{\sqrt{2}}\left[\begin{pmatrix} (D_n - D_n)v_0 \\ -(D_n - D_n)v_0 \end{pmatrix} + \begin{pmatrix} -2J_{n-1}v_0 \\ 2J_{n-1}v_0 \end{pmatrix}\right] = \frac{1}{\sqrt{2}}\begin{pmatrix} -2 \cdot 2^{n-1}v_0 \\ 2 \cdot 2^{n-1}v_0 \end{pmatrix} = -2^n\tilde{v}_{n+1}.$$

By construction, $\langle \tilde{v}_j, \tilde{v}_j \rangle = 1$ for $j = 0, \ldots, n+1$, and since $\langle v_j, v_k \rangle = 0$ for any $j \neq k$, $\tilde{v}_0, \ldots, \tilde{v}_{n+1}$ are orthogonal. By the principle of induction, the assertion is true. □

By Theorem 3.1 and Theorem 3.2,

$$D_n = n2^{n-1}v_0v_0^t - 2^{n-1}(v_1v_1^t + \cdots + v_nv_n^t).$$

This result provides a more efficient way to generate D_n using the $n+1$ eigenvectors. So only $n+1$ vectors of size 2^n have to be stored to construct D_n. In Section 2.1, D_n was generated recursively by D_{n-1}, meaning that to generate D_n, 2^{n-1} vectors of size 2^{n-1} had to be stored.

Next using Theorems 3.1 and 3.2 we can generate the powers of D_n.

Theorem 3.3 *Let k be a positive integer. Then*

$$D_n^k = a(n,k)v_0v_0^t + b(n,k)D_n,$$

where

$$a(n,k) = 2^{k(n-1)}(n^k + (-1)^k n) \quad \textit{and} \quad b(n,k) = (-2^{n-1})^{k-1}.$$

Proof. By Theorem 3.1 and Theorem 3.2,

$$D_n = n2^{n-1}v_0v_0^t - 2^{n-1}(v_1v_1^t + \cdots + v_nv_n^t).$$

Let $L_n = v_1v_1^t + \cdots + v_nv_n^t$. Then $D_n = n2^{n-1}v_0v_0^t - 2^{n-1}L_n$. So $2^{n-1}L_n = n2^{n-1}v_0v_0^t - D_n$. Therefore, $L_n = nv_0v_0^t - 2^{1-n}D_n$. Recall that

$$D_n^k = (n2^{n-1})^k v_0v_0^t + (-2^{n-1})^k L_n.$$

Making the substitution for L_n, yields

$$D_n^k = (n2^{n-1})^k v_0v_0^t + (-2^{n-1})^k[nv_0v_0^t - 2^{1-n}D_n].$$

Regrouping the terms, we get

$$\begin{aligned} D_n^k &= [(n2^{n-1})^k + (-2^{n-1})^k n]v_0v_0^t + (2^{n-1})^{k-1}D_n \\ &= 2^{k(n-1)}(n^k + (-1)^k n)v_0v_0^t + (-2^{n-1})^{k-1}D_n. \end{aligned}$$

The result follows. □

As a consequence of Theorem 3.3, no matter what power k, D_n^k will only have as many distinct values as D_n.

Corollary 3.4 *For every positive integer k, D_n^k has $n+1$ distinct values.*

4 Decomposition of D_n and graph structure of genetic sequences

Since $2^{1-n}D_n$ is a doubly stochastic matrix, it can be decomposed into a convex combination of permutation matrices [10, 19]. We will show that the combination involves only 2^n permutation matrices, which can be defined recursively. The decomposition for $n = 3$ was shown in [8].

Theorem 4.1 *Let D_n be the Hamming distance matrix defined in Section 1. Then*

$$D_n = \sum_{i=1}^{2^n} a_i^n P_i^n,$$

where $(a_1^n, a_2^n, \ldots, a_{2^n}^n)$ with $a_i^n \in \{0, 1, \ldots, n\}$, and P_i^n are permutation matrices determined as follows:
For $n = 1$,

$$(a_1^1, a_2^1) = (0, 1) \quad \textit{and} \quad P_1^1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \quad \textit{and} \quad P_2^1 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

For $n \geq 1$

$$P_j^{n+1} = \begin{pmatrix} P_j^n & 0 \\ 0 & P_j^n \end{pmatrix} \quad \textit{and} \quad P_{j+2^n}^{n+1} = \begin{pmatrix} 0 & P_j^n \\ P_j^n & 0 \end{pmatrix}$$

and

$$(a_1^{n+1}, a_2^{n+1}, \ldots, a_{2^{n+1}}^{n+1}) = (a_1^n, \ldots, a_{2^n}^n, a_1^n, \ldots, a_{2^n}^n) + (\underbrace{0, \ldots, 0}_{2^n}, \underbrace{2, \ldots, 2}_{2^{n-1}}, \underbrace{0, \ldots, 0}_{2^{n-1}}).$$

Moreover, $P_1^n + \ldots + P_{2^{n-1}}^n = J_{n-1}$, and each P_i^n is bisymmetric.

Proof. We prove the result by induction on n, including the additional property that $P_1^n + \ldots + P_{2^{n-1}}^n = J_{n-1}$ and P_i^n is bisymmetric. Take

$$D_1 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} = 0 \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} + 1 \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix},$$

and

$$D_2 = \begin{pmatrix} 0 & 1 & 2 & 1 \\ 1 & 0 & 1 & 2 \\ 2 & 1 & 0 & 1 \\ 1 & 2 & 1 & 0 \end{pmatrix} = 0 \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} + 1 \begin{pmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix} + 2 \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix} + 1 \begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix}.$$

Assume that the scheme is true for n, it will be shown that this is true for $n + 1$. By Theorem 2.1, if

$$D_n = \begin{pmatrix} B_1 & B_2 \\ B_2 & B_1 \end{pmatrix}$$

then

$$D_{n+1} = \begin{pmatrix} B_1 & B_2 & B_1 & B_2 \\ B_2 & B_1 & B_2 & B_1 \\ B_1 & B_2 & B_1 & B_2 \\ B_2 & B_1 & B_2 & B_1 \end{pmatrix} + \begin{pmatrix} 0 & 0 & 2J_{n-1} & 0 \\ 0 & 0 & 0 & 2J_{n-1} \\ 2J_{n-1} & 0 & 0 & 0 \\ 0 & 2J_{n-1} & 0 & 0 \end{pmatrix}.$$

For $j = 1, \ldots, n+1$, define a_j^{n+1} and P_j^{n+1} as in the theorem. Clearly, we have

$$\sum_{j=1}^{2^n} a_j^{n+1} \begin{pmatrix} P_j^n & 0 \\ 0 & P_j^n \end{pmatrix} = \begin{pmatrix} B_1 & B_2 & 0 & 0 \\ B_2 & B_1 & 0 & 0 \\ 0 & 0 & B_1 & B_2 \\ 0 & 0 & B_2 & B_1 \end{pmatrix}.$$

By induction assumption,

$$P_1^{n-1} + \cdots + P_{2^{n-1}}^{n-1} = J_{n-1},$$

we have

$$\sum_{j=1}^{2^{n-1}} P_{j+2^n}^{n+1} = \begin{pmatrix} 0 & 0 & \sum P_j^{n-1} & 0 \\ 0 & 0 & 0 & \sum P_j^{n-1} \\ \sum P_j^{n-1} & 0 & 0 & 0 \\ 0 & \sum P_j^{n-1} & 0 & 0 \end{pmatrix} = \begin{pmatrix} 0 & 0 & J_{n-1} & 0 \\ 0 & 0 & 0 & J_{n-1} \\ J_{n-1} & 0 & 0 & 0 \\ 0 & J_{n-1} & 0 & 0 \end{pmatrix}.$$

It follows that

$$\sum_{j=1}^{2^n} a_{j+2^n}^{n+1} \begin{pmatrix} 0 & P_j^n \\ P_j^n & 0 \end{pmatrix} = \begin{pmatrix} 0 & 0 & 2J_{n-1} + B_1 & B_2 \\ 0 & 0 & B_2 & 2J_{n-1} + B_1 \\ 2J_{n-1} + B_1 & B_2 & 0 & 0 \\ B_2 & 2J_{n-1} + B_1 & 0 & 0 \end{pmatrix}.$$

Thus, D_{n+1} has the asserted combination. It is also easy to check that $P_1^{n+1} + \cdots + P_{2^{n+1}}^{n+1} = J_{n+1}$ using the induction assumption. □

Consider the graph G_n^* using the genetic sequences of C_n as vertices, and two vertices are adjacent if they have a Hamming distance of 1. It is trivial to show that G has a Hamilton circuit, between all the length n nucleotide sequences. Start at position $(1, 1)$ and connect the neighboring entry with an edge and do that for every cell until position $(1, 2^n)$. Then draw and edge from position $(1, 2^n)$ and $(2, 2^n)$, connect the edges in the reverse direction. Repeating this process will connect all 4^n nucleotide sequences of C_n. However we can make a stronger statement: G_n^* has a Hamilton circuit, such that each circuit of the subgraph corresponding to a permutation matrix can be connected to form a Hamilton circuit of all length n nucleotide sequences. The Hamilton circuit which provides a pathway for the genetic code structure [7, 8]. For $n = 2$:

$$P_1 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \sim \begin{pmatrix} CC & 0 & 0 & 0 \\ 0 & CG & 0 & 0 \\ 0 & 0 & GG & 0 \\ 0 & 0 & 0 & GC \end{pmatrix}, \quad P_2 = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix} \sim \begin{pmatrix} 0 & CU & 0 & 0 \\ CA & 0 & 0 & 0 \\ 0 & 0 & 0 & GA \\ 0 & 0 & GU & 0 \end{pmatrix},$$

$$P_3 = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix} \sim \begin{pmatrix} 0 & 0 & UU & 0 \\ 0 & 0 & 0 & UA \\ AA & 0 & 0 & 0 \\ 0 & AU & 0 & 0 \end{pmatrix}, \quad P_4 = \begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix} \sim \begin{pmatrix} 0 & 0 & 0 & UC \\ 0 & 0 & UG & 0 \\ 0 & AG & 0 & 0 \\ AC & 0 & 0 & 0 \end{pmatrix}.$$

So the circuits correspond to these matrices are $CC-CG-GG-GC-CC$, $CU-CA-GA-GU-CU$, $UU-UA-AA-AU-UU$, and $UC-UG-AG-AC-UC$ for P_1, P_2, P_3, P_4, respectively. If the first edge in every circuit is deleted, an edge can be drawn between $CG-CA$, $CU-UU$, $UA-UG$, and $UC-CC$. For $n = 2$ a Hamilton circuit of G_2^* can be constructed as follows:

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

which is a circuit containing all entries of C_2. The pattern to observe is that the first circuit starts by going backwards, and the next circuit runs forwards. This pattern repeats itself until the Hamilton circuit is completed.

Lemma 4.2 *Assume P_i^n is the permutation matrix as defined in Theorem 4.1. If P_i^n has a nonzero entry at position $(1, q_1)$, then the $(2^n, 2^n - q_1 + 1)$ and the $(2^{n-1} + 1, 2^{n-1} - q_{2^{n-1}} + 1)$ entry of P_i^n will also be nonzero.*

Proof. This follows from the bisymmetry of P_i^n. □

Theorem 4.3 *Consider $D_n = \sum_{i=1}^{2^n} a_i^n P_i^n$ as described in Theorem 4.1.*

(a) *Suppose P_i^n has nonzero entries at the $(1, q_1), (2, q_2), \ldots, (2^n, q_{2^n})$ positions where the genetic sequences corresponding to those in C_n are $g_1, g_2, \ldots, g_{2^n}$, of C_n. Then $g_1 - g_2 - \ldots - g_{2^n} - g_1$ is a circuit in G_n^*. In other words, every consecutive pair of sequences in $g_1 - g_2 - \cdots - g_{2^n} - g_1$ differ by one nucleotide.*

(b) *One can combine the circuits in part* (a) *to form a Hamilton circuit in G_n^*.*

Proof. (a) Consider the graph G_n^*. The assertion is clearly true for $n = 2$, by the discussion before Lemma 4.2. So assume the nucleotide sequences corresponding to the nonzero entries of P_i^n can be connected to form a circuit in the graph G_n^*. It will be shown that the nucleotide sequences corresponding to the nonzero entries of P_i^{n+1} and $P_{i+2^n}^{n+1}$, can be connected to form a circuit in the graph G_{n+1}^*.

By induction assumption the nonzero entries of P_i^n corresponds to a circuit in G_n^* denoted as $x_1 - x_2 - \cdots - x_{2^n} - x_1$, where the position of x_1 is $(1, q_1)$, x_2 is $(2, q_2), \ldots$, x_{2_n} is $(2^n, q_{2^n})$. So the nonzero entries at the $(1, q_1), \ldots, (2^n, q_{2^n})$ positions gives rise to a circuit $x_1 - x_2 - \cdots x_{2^n} - x_1$ in G_n^*. By the recursive structure in P_i^{n+1}, the nucleotides corresponding to the nonzero entries of P_i^{n+1} form two disjoint circuits with no common edges, because P_i^n appears as two sub-matrices of P_i^{n+1}. Let the two circuits of P_i^{n+1} be $x_1 - x_2 - \ldots - x_{2^n} - x_1$ and $y_1 - y_2 - \ldots - y_{2^n} - y_1$, respective to the nucleotide sequences. Note that the circuits corresponding positions in the matrix are $(1, q_1) - (2, q_2) - \cdots - (2^n, q_{2^n}) - (1, q_1)$ and $(2^n + 1, r_1) - (2^n + 2, r_2) - \cdots - (2^{n+1}, r_{2_n}) - (2^n + 1, r_1)$, respectively.

By Lemma 4.2, $r_1 = 2^{n+1} - q_{2^n} + 1$, so r_1 and q_{2_n} are equidistant from the vertical center because $r_1 + q_1 = 2^{n+1} + 1$. Also if P_i^{n+1} has a nonzero entry at $(1, q_1)$, then it also has a nonzero entry at $(2^{n+1}, 2^{n+1} - q_1 + 1)$. So since the position corresponding to y_{2^n} is $(2^{n+1}, r_{2^n})$, and $r_{2^n} = 2^{n+1} - q_1 + 1$, so x_1 and y_{2^n} are also equidistant from the center. Since

$$G_n = \{0||a_0, 0||a_1, \ldots, 0||a_{n-1}, 1||a_{n-1}, 1||a_{n-2}, \ldots, 1||a_0\},$$

two Gray codes equidistant from the center only change in the first bit, i.e., the first bit will change from a 0 to a 1 or vise versa. Therefore x_{2^n} and y_1 are adjacent, and y_{2^n} and x_1 are adjacent. So, delete the edges (x_{2^n}, x_1) and (y_{2^n}, y_1), and then connect (x_{2^n}, y_1) and (y_{2^n}, x_1); that will be a circuit in G_{n+1}^*.

(b) Consider the nucleotide sequences corresponding to the nonzero entries in P_i^n, described as in part (a), labeled $x_1^i, x_2^i, \ldots, x_{2^n}^i$. For all $i \leq 2^n - 1$, Delete the edge (x_1^i, x_2^i). As proven in

Corollary 2.2, each neighboring nucleotide has a Hamming distance of 1; also the nucleotide in the last and first column have a Hamming distance of 1. By construction of P_i^n, the i^{th} column of the first row, of P_i^n, is nonzero. Therefore, an edge can be drawn between the nucleotides corresponding to the nonzero entries, in the first row of P_i^n and P_{i+1}^n.

Also by the recursive construction of P_i^n, when i is odd, the nucleotides corresponding to the nonzero entries in the second rows of P_i^n and P_{i+1}^n correspond to neighboring nucleotides. Thus when i is odd, an edge can be drawn between the nucleotides corresponding to the non-zero entries in the second row of P_i^n and P_{i+1}^n; when i is even, draw an edge between the nucleotides corresponding to the nonzero entries in the first row of P_i^n and P_{i+1}^n. Also draw an edge between the nucleotides corresponding to the nonzero entries in the first row of $P_{2^n}^n$ and P_1^n. This creates a Hamilton circuit that is connected via the circuits of part (a). □

5 Further remarks and research

As presented in Section 1, C_n is the genetic code matrix with each cell represented by n-distinct nucleotides, and there is a recursive way to generate C_n.

Evidently, each entry of D_n records the total number of occurrences of U and A in the genetic sequence in the corresponding entry in C_n. Using Theorem 2.1, one can easily extend the construction of D_n and C_n to build a matrix S_n such that each entry is a four tuple recording the number of occurrences of C, U, A, G in the corresponding entry as described in the following.

Theorem 5.1 *Define S_n to be a matrix of size $2^n \times 2^n$, where each cell of S_n is represented by a numerical sequence, (x_C, x_U, x_A, x_G), where x_i is the number of times the i^{th} nucleotide is represented in C_n. Then*

$$S_{n+1} = \begin{pmatrix} (1000)J_n + S_n & (0100)J_n + S_nF_n \\ (0010)J_n + F_nS_n & (0001)J_n + F_nS_nF_n \end{pmatrix}.$$

Note that S_n can also be identified with a 4-tuple of matrices in M_{2^n}, namely,

$$S_n \sim (S_n^C, S_n^U, S_n^A, S_n^G),$$

so that the each entry of S_n^X records the number of times the symbol $X \in \{C, U, A, G\}$ appears in the corresponding entry in C_n. It would be interesting to study the algebraic structure of each S_n^X, and explore the implications to biological study.

As pointed out by one of the referees, information is often stored in computers in matrix form, and matrix methods have been used in the study of many branches of natural and physical sciences such as quantum mechanics. Matrix analysis methods in molecular genetics and bioinformatics have been utilized intensively in the last decade. Our study has revealed new patterns and symmetrical relations in genetic sequences stored in matrix forms. Hopefully, these will inspire new techniques and methods in the study of genetic sequences and bioinformatics. We will say more about implications of our results, together with their limitations and other connections to other study as mentioned by the referees in the following.

Through this paper, we explored information on genetic code and the corresponding Hamming distances that are related to nucleotide strings. This information has been presented in a structurally recursive manner that is easy to generate. An important issue that can be addressed is how to apply the recursive schemes to current biological problems.

There may be interesting implications of the graph structure and Hamilton circuit that could be useful in genetic mutation. Since the two vertices of the graph are adjacent if and only if the codons differ in one position, what effect would changing a codon during RNA transcription have on the corresponding amino acid? For example, if one wanted to compute how many mutations it would take for GCU to mutate into CUC, one could examine all of the pertinent Hamilton paths between the two codons.

Furthermore, we have described an efficient way to generate D_n^k using the eigenstructure. In graph theory, the (i, j) entry of the kth power of the adjacency matrix of a graph counts the number of length k walks from vertex i to vertex j. It would be interesting to find a connection between the entries of D_n^k and the genetic sequences obtained by joining k sequences in C_n.

A referee pointed out that the amino acid code structure using Gray code may be effective in minimizing reading errors, but not for studying mutation errors since mutation errors are random with respect to base position within a codon. Responding to this comment, we think that it is a good idea to include a probability component in the matrix model. For example, as shown in [8, Section 3.2], there is a connection between the entries of the mth power of the matrix D_3 and some simple paths of a certain simple graph with codons as vertices. One may scale the entries of D_3 by some factors to reflect the probabilities connecting two vertices in the simple graph.

Another referee pointed out that in protein evolution one is interested in assigning probabilities to the possible paths between a pair of codons after k steps. Usually, the most parsimonious [3] path is assumed in the construction of phylogenetic trees. A clearly ad hoc assumption, that may not be correct in many cases. To some researchers, this may not be the right approach to phylogenetic trees. Using the entries of the adjacency matrix to evaluate a priori mutation probabilities could be a better choice.

As mentioned in Section 1, there are redundancies in the codons of genetic code, but there is no ambiguity. For example, CCU and CCC both represent Prolic (Pro) acid, but there is no ambiguity so that no codon represents more than one amino acid. There are also start and stop codons. The translation section of genetic code starts with an initiation chain which is called a start codon. Stop codons are identified by the name of a color, and they signal release factors, so there is a mapping that maps the Genetic codons to their amino acids. There are 20 amino acids and 1 start codon, so there is obviously going to be some overlap, which is modeled in this matrix. Note that this is only for $n = 3$ and any multiple of three, since codons are tri-nucleotide sequences. C_n can be mapped from codons to amino acids.

For n=3

$$C_3 = \begin{pmatrix} CCC & CCU & CUU & CUC & UUC & UUU & UCU & UCC \\ CCA & CCG & CUG & CUA & UUA & UUG & UCG & UCA \\ CAA & CAG & CGG & CGA & UGA & UGG & UAG & UAA \\ CAC & CAU & CGU & CGC & UGC & UGU & UAU & UAC \\ AAC & AAU & AGU & AGC & GGC & GGU & GAU & GAC \\ AAA & AAG & AGG & AGA & GGA & GGG & GAG & GAA \\ ACA & ACG & AUG & AUA & GUA & GUG & GCG & GCA \\ ACC & ACU & AUU & AUC & GUC & GUU & GCU & GCC \end{pmatrix}$$

[3] Principle of Parsimony is a minimalist principle, sometimes also referred to as "Ockham's razor," and sates that one should prefer simpler explanation, requiring fewer assumptions over more complex, ad hoc ones. In phylogeny reconstruction, this principle has been applied in two ways. One emphasizes the feature that minimalist principle favors the tree requiring the fewest evolutionary events (such as mutations) to explain the observed data and thus, in some sense, the 'simplest," or an "optimal" description of the the data. A second appeals to the Principal of Parsimony is to assume as little as possible about any underlying model or mechanism for evolution; see for example [17, 20].

But our Amino Acid Matrix (denoted by A_n, where n is a multiple of 3) is as follows

$$A_3 = \begin{pmatrix} Pro & Pro & Leu & Leu & Phe & Phe & Ser & Ser \\ Pro & Pro & Leu & Leu & Leu & Leu & Ser & Ser \\ Gln & Gln & Arg & Arg & OPAL & Trp & AMBER & OCHRE \\ His & His & Arg & Arg & Cys & Cys & Tyr & Tyr \\ Asn & Asn & Ser & Ser & Gly & Gly & Asp & Asp \\ Lys & Lys & Arg & Arg & Gly & Gly & Glu & Glu \\ Thr & Thr & MET(START) & Ile & Val & Val & Ala & Ala \\ Thr & Thr & Ile & Ile & Val & Val & Ala & Ala \end{pmatrix}$$

Note that in A_3, $MET, OPAL, AMBER$ and $OCHRE$, are the start and stop codons as mentioned in the previous paragraph. Note that the matrix A_3 is for the so called the Standard code, which is one of many dialects of the genetic code. All dialects are presented in the NCBIs site http://www.ncbi.nlmnih.gov/Taxonomy/Utils/wprintgc.cgi. The matrix approach on the basis of the Hamming distance should be applied for all dialects in future. It would be interesting to encode and study the matrix A_n.

We close the paper by presenting some additional inspiring remarks of two referees.

A referee has the following reservation for our work as follows. "The work in this paper develops matrix machinery to represent and compare all possible nucleotide sequences of a given length, but this does not lead to any new biological insights. From a practical standpoint this representation is too abstract and general to be useful to a biologist. For example, for a 100 nucleotide sequence, which is tiny, the matrix machinery considers 2^{100} (*note by the authors: it should actually be* 4^{100}) possible sequences, or 10^{30} sequences. In the entirety of Genbank, there are only on the order of 10^8 sequences of all lengths. In conclusion, the referee thinks that the suggested potential applications we mentioned can be accomplished by less cumbersome machinery, likely by existing software."

We certainly agree with the referee that there is much room for improvement of our results. Nevertheless, our results do give efficient way to store and manipulate the data. For example, not only can we obtain an efficient algorithm to generate the $2^n \times 2^n$ matrices C_n and D_n by our results in Section 2, we can use the results in Section 3 to represent the $2^n \times 2^n$ matrix D_n in terms of the eigenvalues $n2^{n-1}$ and -2^{n-1} (multiplicity n) together with their eigenvectors $n+1$ vectors $v_0, \ldots, v_n \in \mathbf{R}^{2^n}$, which requires the storage of $2 + (n+1)2^{n+1}$ numbers. Moreover, the power of D_n can be expressed as a combination of $v_0 v_0^t$ and D_n, that requires hardly any extra memory to compute and store. Moreover, results in Section 4 provide systematic ways to decompose the matrices D_n into sum of permutation matrices corresponding to Hamiltonian graph structure that may have implications to the study of mutations. Even if our decompositions may not be most effective in studying patterns arising in biology applications, the general ideas and techniques may be modified and adapted to study important problems.

To a certain extent, the following comment of another referee may help put our work in perspective. "The modern situation in the theoretic field of genetic informatics can be characterized by the following statement by famous researches from GenBank: 'What will we have when these genomic sequences are determined? What do we have now in the 10 million nucleotide of sequence data determined to date? We are in the position of Johann Kepler when he first began looking for patterns in the volumes of data that Tycho Brahe had spent his life accumulating. We have the program that runs the cellular machinery, but we know very little about how to read it. Bench biologists, by experiment and by close association with the data, have found meaningful patterns. Theoreticians, by careful reasoning and use of collections of data, have found others, but we still

understand frustratingly little;' see [2]. Kepler is mentioned here not without reason. The history of science shows the importance of cognitive forms of presentation of phenomenological data to find regularities or laws in this phenomenology. The work by Kepler is the classical example of an important meaning of a cognitive form of presentation of phenomenological data. He did not make his own astronomic observations, but he found the cognitive form of presentation in the huge astronomic data from the collection of Tycho Brahe. This discovered form, which was connected to the general idea of movements along ellipses, allowed him to formulate the famous Kepler's laws of planetary movements relative to the Sun. Owing to this cognitive form, Kepler and Newton have led us to the law of Newtonian attraction. A discovery of such a cognitive form of presentation in the case of the phenomenology of genetic code systems is a modern challenge, which arises from the very beginning in the course of attempts to find regularities among a huge number of genetic data and to create a relevant theory. Matrix genetics proposes a new cognitive form of presentation of phenomenological data in the field of genetic informatics. This cognitive matrix form gives new tools to analyze and to model ensembles of the genetic code as well. It paves the way for a worthy attempt at answering the mentioned challenges. This article belongs to this actual direction and proposes interesting improvements of relevant mathematical apparatus. Matrix genetics gives new results which reveal new branches of biological and bio-mathematical researches; for example see [9, 14]."

As pointed out by the editor, the work in [9, 14] is more theoretical and provides a methodology that may be relevant in the future. We believe that our paper belongs to the same category.

Acknowledgment

The authors would like to thank Professor M. He for drawing their attention to the interesting topic, sending them the reprints of [7, 8] and some helpful comments. They also thank the three referees for their careful reading of the paper, and their valuable suggestions including footnote (3), which are incorporated in the last section of the paper.

References

[1] C. Alff-Steinberger, The genetic code and error transmission, Proc. Natl. Acad. Sci. USA 64 (1969), 584-591.

[2] J. Fickett and Chr. Burks, Development of a database for nucleotide sequences. - In: M.S.Waterman (Ed.), Mathematical Methods in DNA Sequences, pp.1-34. Florida: CRC Press, 1989.

[3] S.J. Freeland, T. Wu and N. Keulmann, The case for an error minimizing genetic Code, Origins of Life and Evolution of Biospheres 33 (2003), 457-77.

[4] M. A. Gates, A simple way to look at DNA, J. Theor. Biol. 119 (1986), 319328.

[5] A.L. Goldberg and R.E. Wittes, Genetic code: aspects of organisation, Science 153 (1966), 420-424.

[6] D. Haig and L.D. Hurst, A quantitative measure of error minimization in the genetic code, J. Mol. Evol. 22 (1991), 412-417.

[7] M.X. He, Genetic code, Attributive mappings and stochastic matrices, Bull. Math. Biology 66 (2004), 965-973.

[8] M.X. He, S.V. Petoukhov and P.E. Ricci, Genetic code, Hamming distance and stochastic matrices, Bull. Math. Biology 66 (2004), 1405-1421.

[9] M.X. He and S.V. Petoukhov, Harmony of living nature, symmetries of genetic systems and matrix genetics, International J. of Integrative Biology 1 (2007), 41-43.

[10] R.A. Horn and C.R. Johnson, Matrix Analysis, Cambridge University Press, New York, 1985.

[11] M.A. Jimenéz-Monteno, C.R. Mora-Basenez and T. Poechel, The Hypercube Structure of Genetic Code, BioSystems, 39 (1996), 117-125.

[12] P. M. Leong and S. Morgenthaler, Random walk and gap plots of DNA sequences, Comput. Applic. Biosc. 11 (1995), 503507.

[13] A. Nandy, A new graphical representation and analysis of DNA sequence structure: I. Methodology and application to globin genes, Curr. Sci. 66 (1994), 309314.

[14] S. Petoukhov, Matrix genetics, algebras of the genetic code, noise-immunity, Moscow, RCD, 2008.

http://www.geocities.com/symmetrion/Matrix_genetics/matrix_genetics.html)

[15] J.H. Reif, Alternative Computational Models: A Comparison of Biomolecular and Quantum Computation, an invited paper at the 18th International Conference on Foundations of Software Technology and Theoretical Computer Science (FST&TCS98), 1998. Preprint can be found at http://www.cs.duke.edu/˜reif/paper/altcomp.ps.

[16] T.M. Sonneborn, Degeneracy of the genetic code: extent, nature and genetic implications, Academic Press, New York, 1965.

[17] M. Steel and D. Penny, Parsimony, likelihood, and the role of models in molecular phylogenetics, Mol. Bio. & Evol. 17 (2000), 839-850.

[18] R. Swanson, A Unifying Concept for The Amino Acid Code. Bull. Math. Biology. 46 (1984), 187-203.

[19] A. Tucker, Applied Combinatorics (5^{th} ed.), John Wiley & Sons, New York, 2007.

[20] Z. Yang, Phylogenetic Analysis using parsimony and likelihood methods, J. Mol. Evol. 42 (1996), 294-307.