Genetic Code

A Matrix and Combinatoric Approach

Tanner Crowder

April 5, 2008

Introduction

- Genetic and Gray Code
- Matrices Relating to Genetic Code

Eigenstructure of D_n

- Eigenvalues
- Eigenvectors

The Genetic Code Matrix C_n

- Basic Structures of C_n
- Hypercube Structure

Acknowledgments

Thanks to Dr. Chi-Kwong Li, for advising me on my honors thesis and his dedication to this work.
I would also like to thank the Professors at William and Mary that have made CSUMS possible.

Genetic code

- Genetic Code is the set of rules by which information is encoded in DNA/RNA that is translated into amino acid sequences by living cells.

Genetic code

- Genetic Code is the set of rules by which information is encoded in DNA/RNA that is translated into amino acid sequences by living cells.
- Nucleotides are the basis for encoding which are labeled $\{C, U, A, G\}$

Genetic code

- Genetic Code is the set of rules by which information is encoded in DNA/RNA that is translated into amino acid sequences by living cells.
- Nucleotides are the basis for encoding which are labeled $\{C, U, A, G\}$
- A genetic code map is $g : C' \longrightarrow A'$, where $C' = (\{x_1x_2x_3\}) : x_i \in R = \{A, C, G, U\}$, where C' is the set of codons, and A' are the amino acids and termination codons.

Genetic code

- Genetic Code is the set of rules by which information is encoded in DNA/RNA that is translated into amino acid sequences by living cells.
- Nucleotides are the basis for encoding which are labeled $\{C, U, A, G\}$
- A genetic code map is $g : C' \rightarrow A'$, where $C' = (\{x_1 x_2 x_3\}) : x_i \in R = \{A, C, G, U\}$, where C' is the set of codons, and A' are the amino acids and termination codons.
- The focus of this study is building matrices for different length nucleotide sequences, and how to represent the sequences more efficiently

Gray Code

- A Gray-code representation of the nucleotides was proposed by Swanson (Swanson, 1984)
- Gray code is a set of binary sequences with the property that two consecutive number only differ by one position

Gray Code

- A Gray-code representation of the nucleotides was proposed by Swanson (Swanson, 1984)
- Gray code is a set of binary sequences with the property that two consecutive number only differ by one position
- Example: In classical binary three and four are 011 and 100 respectively. In Gray Code, the 3 bit representations for three and four are 011 and 010, respectively.

Gray Code

- A Gray-code representation of the nucleotides was proposed by Swanson (Swanson, 1984)
- Gray code is a set of binary sequences with the property that two consecutive number only differ by one position
- Example: In classical binary three and four are 011 and 100 respectively. In Gray Code, the 3 bit representations for three and four are 011 and 010, respectively.
- In genetic transcription a mismatch in genetic coding segments will reduce the degree of mutation

Gray Code

- Let G_n, be all the Gray code sequences of length n; G_n can be generated by a recursive algorithm.

Gray Code

- Let G_n, be all the Gray code sequences of length n; G_n can be generated by a recursive algorithm.
- $G_n = \{0||a_0, 0||a_1, \ldots, 0||a_{n-1}, 1||a_{n-1}, 1||a_{n-1}, \ldots, 1||a_0\}$, where $a_i \in G_{n-1}$
- Example: $G_1 = \{0, 1\}$, then $G_2 = \{0||0, 0||1, 1||1, 1||0\} = \{00, 01, 11, 10\}$

Gray Code

- Let G_n, be all the Gray code sequences of length n; G_n can be generated by a recursive algorithm.
- $G_n = \{0||a_0, 0||a_1, \ldots, 0||a_{n-1}, 1||a_{n-1}, 1||a_{n-1}, \ldots, 1||a_0\}$, where $a_i \in G_{n-1}$
- Example: $G_1 = \{0, 1\}$, then $G_2 = \{0||0, 0||1, 1||1, 1||0\} = \{00, 01, 11, 10\}$
- Clearly since $|G_n|$ doubles in size from $|G_{n-1}|$ and G_1 only has 2 entries, $|G_n| = 2^n$

Gray Code

- Let G_n, be all the Gray code sequences of length n; G_n can be generated by a recursive algorithm.
- $G_n = \{0||a_0, 0||a_1, \ldots, 0||a_{n-1}, 1||a_{n-1}, 1||a_{n-1}, \ldots, 1||a_0\}$, where $a_i \in G_{n-1}$
- Example: $G_1 = \{0, 1\}$, then $G_2 = \{0||0, 0||1, 1||1, 1||0\} = \{00, 01, 11, 10\}$
- Clearly since $|G_n|$ doubles in size from $|G_{n-1}|$ and G_1 only has 2 entries, $|G_n| = 2^n$
- Using two bit Gray code construction $C \sim \binom{0}{0}$, $U \sim \binom{1}{0}$, $G \sim \binom{1}{1}$, and $A \sim \binom{0}{1}$.

Genetic Code Matrix

- Define C_n as the genetic code matrix as the matrix that contains all length n nucleotide sequences

Genetic Code Matrix

- Define C_n as the genetic code matrix as the matrix that contains all length n nucleotide sequences
- $C_1 \sim \begin{array}{c c} & \begin{array}{cc} 0 & 1 \end{array} \\ \begin{array}{c} 0 \\ 1 \end{array} & \begin{pmatrix} \binom{0}{0} & \binom{1}{0} \\ \binom{0}{1} & \binom{1}{1} \end{pmatrix} \end{array}$

Genetic Code Matrix

- Define C_n as the genetic code matrix as the matrix that contains all length n nucleotide sequences
- $C_1 \sim \begin{array}{cc} & \begin{matrix} 0 & 1 \end{matrix} \\ \begin{matrix} 0 \\ 1 \end{matrix} & \begin{pmatrix} \binom{0}{0} & \binom{1}{0} \\ \binom{0}{1} & \binom{1}{1} \end{pmatrix} \end{array}$
- $C_1 = \begin{pmatrix} C & U \\ A & G \end{pmatrix}$

- $$C_2 \sim \begin{array}{c|cccc} & 00 & 01 & 11 & 10 \\ 00 & \binom{00}{00} & \binom{01}{00} & \binom{11}{00} & \binom{10}{00} \\ 01 & \binom{00}{01} & \binom{01}{01} & \binom{11}{01} & \binom{10}{01} \\ 11 & \binom{00}{11} & \binom{01}{11} & \binom{11}{11} & \binom{10}{11} \\ 10 & \binom{00}{10} & \binom{01}{10} & \binom{11}{10} & \binom{10}{10} \end{array}$$

- $$C_2 \sim \begin{array}{cc} & \begin{array}{cccc} 00 & 01 & 11 & 10 \end{array} \\ \begin{array}{c} 00 \\ 01 \\ 11 \\ 10 \end{array} & \begin{pmatrix} \binom{00}{00} & \binom{01}{00} & \binom{11}{00} & \binom{10}{00} \\ \binom{00}{01} & \binom{01}{01} & \binom{11}{01} & \binom{10}{01} \\ \binom{00}{11} & \binom{01}{11} & \binom{11}{11} & \binom{10}{11} \\ \binom{00}{10} & \binom{01}{10} & \binom{11}{10} & \binom{10}{10} \end{pmatrix} \end{array}$$

- $$C_2 = \begin{pmatrix} CC & CU & UU & UC \\ CA & CG & UG & UA \\ AA & AG & GG & GA \\ AC & AU & GU & GC \end{pmatrix}$$

Hamming Distances

- The hamming distance is a measure of how different are two strings of the same length

Hamming Distances

- The hamming distance is a measure of how different are two strings of the same length
- For example the codon $CAG \sim \binom{001}{011}$ has a hamming distance of 1, because the second position is different.

Hamming Distances

- The hamming distance is a measure of how different are two strings of the same length
- For example the codon $CAG \sim \binom{001}{011}$ has a hamming distance of 1, because the second position is different.
- Define D_n as the hamming distance matrix that computes the hamming distance between the entries of the cells of the genetic code matrix.

D_1 and D_2

- Recall

$$C_1 \sim \begin{array}{cc} & \begin{array}{cc} 0 & 1 \end{array} \\ \begin{array}{c} 0 \\ 1 \end{array} & \begin{pmatrix} \binom{0}{0} & \binom{1}{0} \\ \binom{0}{1} & \binom{1}{1} \end{pmatrix} \end{array} \quad \text{so } D_1 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$

D_1 and D_2

- Recall

$$C_1 \sim \begin{array}{c c} & \begin{array}{cc} 0 & 1 \end{array} \\ \begin{array}{c} 0 \\ 1 \end{array} & \begin{pmatrix} \binom{0}{0} & \binom{1}{0} \\ \binom{0}{1} & \binom{1}{1} \end{pmatrix} \end{array} \quad \text{so } D_1 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$

- And

$$C_2 \sim \begin{array}{c c} & \begin{array}{cccc} 00 & 01 & 11 & 10 \end{array} \\ \begin{array}{c} 00 \\ 01 \\ 11 \\ 10 \end{array} & \begin{pmatrix} \binom{00}{00} & \binom{01}{00} & \binom{11}{00} & \binom{10}{00} \\ \binom{00}{01} & \binom{01}{01} & \binom{11}{01} & \binom{10}{01} \\ \binom{00}{11} & \binom{01}{11} & \binom{11}{11} & \binom{10}{11} \\ \binom{00}{10} & \binom{01}{10} & \binom{11}{10} & \binom{10}{10} \end{pmatrix} \end{array} \quad \text{so } D_2 = \begin{pmatrix} 0 & 1 & 2 & 1 \\ 1 & 0 & 1 & 2 \\ 2 & 1 & 0 & 1 \\ 1 & 2 & 1 & 0 \end{pmatrix}$$

Basic Properties of D_n

Theorem

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

Basic Properties of D_n

Theorem

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

(ii) *The matrix D_n is doubly stochastic and symmetric.*

Basic Properties of D_n

Theorem

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

(ii) *The matrix D_n is doubly stochastic and symmetric.*

(iii) *D_n is centrally embedded in D_{n+1}*

Recursion in D_n

Theorem

Let $D_n = \begin{pmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{pmatrix}$ *where* B_{ij} *is a* $2^{n-1} \times 2^{n-1}$ *sub matrix.*

Then

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

Recursion in D_n

Theorem

Let $D_n = \begin{pmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{pmatrix}$ *where* B_{ij} *is a* $2^{n-1} \times 2^{n-1}$ *sub matrix.*

Then

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

- Notice that if D_{n+1} is written in the 4×4 block structure, D_n appears centrally embedded as a 2×2 block.

Recursion in D_n

Theorem

Let $D_n = \begin{pmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{pmatrix}$ *where* B_{ij} *is a* $2^{n-1} \times 2^{n-1}$ *sub matrix.*

Then

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

- Notice that if D_{n+1} is written in the 4×4 block structure, D_n appears centrally embedded as a 2×2 block.
- D_n stores information about C_n, however reduces the amount of information stored by a factor of n.

Eigenvalues of D_n

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

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

Eigenvalues of D_n

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

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

- This is fortunate because unlike everything else so far, this is not recursive

Eigenvalues of D_n

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

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

- This is fortunate because unlike everything else so far, this is not recursive
- Notice the first eigenvalue is the column row sum.

Eigenvectors of D_n

- Recall $D_1 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$
- It is easy to see that a set of orthonormal eigenvectors are $v_0 = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ 1 \end{pmatrix}$ and $v_1 = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ -1 \end{pmatrix}$

Eigenvectors of D_n

- Also recall that $D_2 = \begin{pmatrix} 0 & 1 & 2 & 1 \\ 1 & 0 & 1 & 2 \\ 2 & 1 & 0 & 1 \\ 1 & 2 & 1 & 0 \end{pmatrix}$
- A set of orthonormal eigenvectors are $v_0 = \frac{1}{2}\begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}$, $v_1 = \frac{1}{2}\begin{pmatrix} 1 \\ -1 \\ 1 \\ -1 \end{pmatrix}$, $v_2 = \frac{1}{2}\begin{pmatrix} 1 \\ -1 \\ -1 \\ 1 \end{pmatrix}$], and $v_3 = \frac{1}{2}\begin{pmatrix} 1 \\ 1 \\ -1 \\ -1 \end{pmatrix}$

Eigenvectors of D_n

- There is recursion in the eigenvectors of D_n

Eigenvectors of D_n

- There is recursion in the eigenvectors of D_n
- $\tilde{v}_j = \frac{1}{\sqrt{2}}\begin{pmatrix} v_j \\ v_j \end{pmatrix}$ for $j = 0, \ldots, n-1$, $\tilde{v}_n = \frac{1}{\sqrt{2}}\begin{pmatrix} v_n \\ -v_n \end{pmatrix}$ and $\tilde{v}_{n+1} = \frac{1}{\sqrt{2}}\begin{pmatrix} v_0 \\ -v_0 \end{pmatrix}$

Eigenvectors of D_n

- There is recursion in the eigenvectors of D_n
- $\tilde{v}_j = \frac{1}{\sqrt{2}} \begin{pmatrix} v_j \\ v_j \end{pmatrix}$ for $j = 0, \ldots, n-1$, $\tilde{v}_n = \frac{1}{\sqrt{2}} \begin{pmatrix} v_n \\ -v_n \end{pmatrix}$ and $\tilde{v}_{n+1} = \frac{1}{\sqrt{2}} \begin{pmatrix} v_0 \\ -v_0 \end{pmatrix}$
- These form an orthonormal set of eigenvectors of D_{n+1} corresponding to the nonzero eigenvalues.

Eigenvectors of D_n

- There is recursion in the eigenvectors of D_n
- $\tilde{v}_j = \frac{1}{\sqrt{2}} \begin{pmatrix} v_j \\ v_j \end{pmatrix}$ for $j = 0, \ldots, n-1$, $\tilde{v}_n = \frac{1}{\sqrt{2}} \begin{pmatrix} v_n \\ -v_n \end{pmatrix}$ and $\tilde{v}_{n+1} = \frac{1}{\sqrt{2}} \begin{pmatrix} v_0 \\ -v_0 \end{pmatrix}$
- These form an orthonormal set of eigenvectors of D_{n+1} corresponding to the nonzero eigenvalues.
- This lets us write the powers of D_n^k, in terms of the knowing nothing other than D_n, n and k

The Basics of C_n

- The genetic code-base matrix C_n is a $2^n \times 2^n$ matrix with RNA bases of length n. Each two neighboring entries of genetic code, in both directions differs by exactly one base.
- The genetic code matrix can be defined recursively

The Basics of C_n

- The genetic code-base matrix C_n is a $2^n \times 2^n$ matrix with RNA bases of length n. Each two neighboring entries of genetic code, in both directions differs by exactly one base.
- The genetic code matrix can be defined recursively
- If C_n is the genetic code matrix then

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

- Note that F_n is a matrix that has 1's on it's off diagonal

Number of Nucleotides per Cell

- This leads to counting the number of nucleotides per cell, which would store double the information of the hamming distance matrix as a 4-tuple

Number of Nucleotides per Cell

- This leads to counting the number of nucleotides per cell, which would store double the information of the hamming distance matrix as a 4-tuple
- The 4-tuple would be (x_C, x_U, x_A, x_G), where $x_i =$ number of that nucleotide per cell.

Number of Nucleotides per Cell

- This leads to counting the number of nucleotides per cell, which would store double the information of the hamming distance matrix as a 4-tuple
- The 4-tuple would be (x_C, x_U, x_A, x_G), where $x_i =$ number of that nucleotide per cell.
- Label the matrix S_n to count the number of nucleotides per cell. Then

$$S_{n+1} = \begin{pmatrix} (1000) + S_n & (0100) + S_n F_n \\ (0010) + F_n S_n & (0001) + F_n S_n F_n \end{pmatrix}$$

Decomposition of D_n and Hypercube

- Because D_n is doubly stochastic it is decomposable into a convex combination of permutation matrices that have a leading coefficient of $\{0, 1, \ldots, n\}$.

Decomposition of D_n and Hypercube

- Because D_n is doubly stochastic it is decomposable into a convex combination of permutation matrices that have a leading coefficient of $\{0, 1, \ldots, n\}$.
- The structure can be built recursively as well

Decomposition of D_n and Hypercube

- Because D_n is doubly stochastic it is decomposable into a convex combination of permutation matrices that have a leading coefficient of $\{0, 1, \ldots, n\}$.
- The structure can be built recursively as well
- Each permutation matrix can be as a vertex of a hypercube and within that vertex there is a subcube

Decomposition of D_n and Hypercube

- Because D_n is doubly stochastic it is decomposable into a convex combination of permutation matrices that have a leading coefficient of $\{0, 1, \ldots, n\}$.
- The structure can be built recursively as well
- Each permutation matrix can be as a vertex of a hypercube and within that vertex there is a subcube
- There is also a Hamilton circuit between the all nucleotide sequences, where two nucleotide sequences are adjacent if and only if they differ by exactly one position.

Decomposition of D_n and Hypercube

- Because D_n is doubly stochastic it is decomposable into a convex combination of permutation matrices that have a leading coefficient of $\{0, 1, \ldots, n\}$.
- The structure can be built recursively as well
- Each permutation matrix can be as a vertex of a hypercube and within that vertex there is a subcube
- There is also a Hamilton circuit between the all nucleotide sequences, where two nucleotide sequences are adjacent if and only if they differ by exactly one position.
- This may be promising in the study of mutations in genetic code.

Further Research

- It would be useful to get the most information of C_n, or S_n and D_n, without having to display an exponential amount of information.

Further Research

- It would be useful to get the most information of C_n, or S_n and D_n, without having to display an exponential amount of information.
- Information is lost with D_n and S_n, i.e. order of nucleotides

Further Research

- It would be useful to get the most information of C_n, or S_n and D_n, without having to display an exponential amount of information.
- Information is lost with D_n and S_n, i.e. order of nucleotides
- Eventually it would be useful to construct matrices that were polynomial in size