A New Model of Permutation the Pieces of Nucleotides in DNA Sequences Using the Action of Dihedral Group and Graph Theory

Husein Hadi Abbas¹, Hussein Salman Mohammed Hussein²

Department of Mathematics, Faculty of Education for Girls, University of Kufa, Iraq

Abstract: In this paper, we give new model of genetic algorithm using the action of largest subgroup of dihedral Group $D_n$ $n = 3^m$, $m \in \mathbb{N}$ $m \geq 2$ bipartite graph, and a Markov basis for $n^2 - 2n \times 3 \times n^2$ contingency tables with fixed two dimensional marginals such that Bis $H$-invariant.

Keywords: Computational algebraic statistics, sufficient statistics, linear transformation, connected graph, bipartite graph, dihedral group, algebra statistic, Markov basis.

1. Introduction

Let $I$ be a finite set $n = |I|$ elements, we call an element of $I$ a cell and denoted by $i \in I$. $i$ is often multi-index $i = i_1 \ldots i_m$. A non-negative integer $x_i \in \mathbb{N}$ denotes the frequency of a cell $i$. The set of frequencies is called a contingency table and denoted as $x = \{x_i\}_{i \in I} \in \mathbb{N}^I$ with an appropriate ordering of the cells, we treat a contingency table $x = \{x_i\}_{i \in I} \in \mathbb{N}^I$ as a $n$-dimensional column vector of non-negative integers. Not that a contingency table can also be considered as a function from $I$ to $\mathbb{N}$ defined as $i \mapsto x_i$. The $L_1$-norm of $x \in \mathbb{N}^I$ is called the samplesize and denoted as $|x| = \sum_{i \in I} x_i$. We will denote $\mathcal{Z}$ be the set of integer numbers, also we denote to the $a_j \in \mathbb{Z}^v$, $j = 1, \ldots, v$, as fixed column vectors consisting of integers. A $v$-dimensional column vector $t = (t_1, \ldots, t_j) \in \mathbb{Z}^v$ as $t_j = a'_j x$, $j = 1, \ldots, v$. Here $t_j$ denotes the transpose of a vector or matrix. We also define a $v \times p$ matrix $A$, with its $j$-row being $a'_j$ given by $A = [a'_1 \ldots a'_v]$, and if $t = Ax$ is a $v$-dimensional column vector, we define the set $T = \{t: t = Ax, x \in \mathbb{N}^I\} = \{A x \in \mathbb{N}^I\}$. In typical situations of a statistical theory, $t$ is sufficient statistic for the nuisance parameter. The set of $x$'s for a given $t$, $A^{-1}[t] = \{x \in \mathbb{N}^I: Ax = t\}$ ($t$-fibers), is considered for performing similar tests, for the case of the independence model of two-way contingency tables, for example, $t$ is the row sums and column sums of $x$, and $A^{-1}[t]$ is the set of all the same row sums and column sums to $t$. The set of $t$-fibers gives a decomposition of $\mathbb{N}^I$. An important observation is that $t$-fiber depends on given only through its kernel, $\ker(A)$. For different $A$'s with the same kernel, the set of $t$-fibers are the same. In fact, if we define $x_1 \sim x_2 \Leftrightarrow x_1 - x_2 \in \ker(A)$ this relation is an equivalence relation and $\mathbb{N}^I$ is partitioned into disjoint equivalence classes. The set of $t$-fibers is simply the set of these equivalence classes. Furthermore, $t$ may be considered as labels of these equivalence classes. $A$ $n$-dimensional column vector of integers $z = \{z_i\}_{i \in I} \in \mathbb{Z}^n$ is called a move if it is in the kernel of $A$, i.e. $Az = 0 [10]$.

For a move $z$, the positive part $z^+ = \{z^+_i\}_{i \in I}$ and the negative part $z^- = \{z^-_i\}_{i \in I}$ are defined by $z^+_i = \max(z_i, 0)$, $z^-_i = \max(-z_i, 0)$, respectively. Then $z = z^+ - z^-$ and $z^+, z^- \in \mathbb{N}^n z^+, z^- \in \mathbb{N}^n$. Moreover, $z^+$ and $z^-$ are in the same $t$-fiber, i.e., $z^+, z^- \in A^{-1}[t]$ for $t = Az^+ = Az^-$. We define the degree of $z$ as the sample size of $z^+$ or $(z^-)$ and denote it by $\deg(z) = |z^+| = |z^-|$. In the following we denote the set of moves (for a given $A$) by $M = M_q = \mathbb{Z}^n \cap \ker(A) [1]$. Let $G$ be a group and $W$ be a set. A left action of $G$ on $W$ is a function from $G \times W$ to $W$, usually denoted by $(g, w) \mapsto w \in W$ such that $g(hw) = (gh)w$ and $ew = w$ for all $g, h \in G$ and $w \in W$. Where $e$ is the identity element of $G$. We also say that $G$ acts on $W$ on the left.

Let a group $G$ act on a set $W$, and $U \subseteq W$, $G_{(U)} = \{g: gu = u, \forall u \in U\}$ is called the pointwise stabilizer of $U$. Let a group $G$ acts on a set $W$, $U \subseteq W$, and $GU = \{gu: u \in U, g \in G\}$. We call $U$ invariant under $G$ (or $G$-invariant) if $GU = U [9]$.

Let $A: \mathbb{Z}^v \rightarrow \mathbb{Z}^p$ be a linear transformation, $t \in \mathbb{Z}^p$, and $A^{-1}[t]$ be the set of $t$-fibers, and let $B \subseteq \ker(A)$, then we define $A^{-1}[t]_B$ be the graph with vertex set $A^{-1}[t]$ and $u \sim v$ an edge if and only if $u - v \in \pm B [12]$. Let $A^{-1}[t] = \{x \in \mathbb{N}^I: Ax = t\}$. A set of finite moves $B$ is called Markov basis if for all $t$, $A^{-1}[t]$ constitutes one $B$ equivalence class, if $B \subseteq \ker(A)$ is a set such that $A^{-1}[t]_B$ is connected for all $t$, then $B$ is a Markov basis for $A [1]$.
If a group $G$ act on $A^{-1}[t]$ on the left, $B$ is a Markov basis, and $G(B) = \{ g(x): x \in B, g \in G \}$, $B$ is called invariant under $G$ (or $G$-invariant) if $G(B) = B$. We will denote to the polynomials in the $p$ indeterminates (polynomial variables) $p_1, p_2, \ldots, p_p$ over the complex field $\mathbb{C}$ by either $G[p_1, p_2, \ldots, p_p]$ or $G[P] = (p_1, p_2, \ldots, p_p)$. Let $A:Z^n \rightarrow Z^2$ be a linear transformation, the toric ideal $I_A$ is the ideal $<P - P^t; u, v \in \mathbb{N}^n, A(u) = A(v) \geq G[p_1, p_2, \ldots, p_p]$, where $P := p_1^{u_1} \cdot \ldots \cdot p_p^{u_p}$ [11].

In this paper, we use the Markov basis $B$ and action of the subgroup $H$ of dihedral group $G_{2m}$ on these contingency tables to give a new model of permutation the pieces of nucleotides in DNA sequences.

2. Preliminaries

In this section, we review some basic definitions and notations of dihedral group, connected graph, bipartite graph, moves, Markov basis, and toric ideals that we need in our work.

Definition 1 (see [12]). Let $n$ be a positive integer greater than or equal 3. The group of all symmetries of the regular polygon with $n$ sides, including both rotations and reflections, is called the dihedral group and denoted by $D_n$. If we center the regular polygon at origin then the elements of the dihedral group acts as linear transformation of the plane. Let us represent the elements of $D_n$ as matrix, with composition multiplication. Dihedral groups are among the simplest examples of finite groups and they play an important role in group theory, geometry, and chemistry.

The set of rotations is generated by $r$, counterclockwise rotation with angle $2\pi/n$ of order $n$, and the set of reflections is of order 2 and every element $sr^j$ generates $\{ r, sr, s^2r, \ldots, sr^{n-1} \}$. In general, we can write $D_n$ as

$$D_n := \{ sr^k: 0 \leq k \leq n - 1, 0 \leq j \leq 1 \}$$

which has the following properties:

- $r^2 = 1, sr^j s = r^{-j}$
- $sr^k = 1, for all 0 \leq k \leq n - 1$.

The composition of two elements of $D_n$ is given by $r^{i} r^{j} = r^{i+j}, s r^i = r^{-i}, s r^j = r^{j+1}, s^2 r^i = r^{i+1}, s^2 r^j = r^{j-1}$.

Remark 2 (see [12]). If we label the vertices (of the regular $n$-gon) 1 to $n$ in a counterclockwise direction around $n$-gon then the elements of $D_n$ can be written as permutations of vertices, let $r$ be a counterclockwise rotation, and let $s$ be the reflection of the $n$-gon about an axis through the center and vertex 1, as indicated in below. The element $r$ generates the cyclic group of order $n$ $C_n$ which is a normal cyclic subgroup of $D_n$. In all cases, addition and subtraction should be performed using modular arithmetic with modulus $n$.

Elements of $C_n$

Any symmetry will fix the origin and is determined by the image of two adjacent vertices, say 1 and 2. The vertex 1 can be taken to any of $n$ vertices and then the vertex 2 must be taken to one of the two vertices adjacent to the image of 1. Hence, $D_n$ is a non-abelian group of order $2n$ generated by $r$ and $s$.

Now, we give some concepts about the action of a group on a set that we use later.

Definition 3 (see [3]). A graph $G$ is connected if for every pair of distinct vertices $u, v \in V(G)$, where $V(G)$ be the set of vertices of the graph $G$, the graph $G$ has a $u, v$-path. Otherwise, we say the graph is disconnected.

Definition 4 (see [3]). A graph $G$ is a bipartite graph if there are $X, Y \subseteq V(G)$ meeting the following conditions:

1. $V(G) = X \cup Y$
2. $X \cap Y = \emptyset$
3. $G[X]$ and $G[Y]$ are both null graphs, where $G[X]$ and $G[Y]$ are subgraphs of the graph $G$ induced by the set of vertices $X, Y \subseteq V(G)$ respectively.

Theorem 5 (see [3]). For a graph $G$ the following statements are equivalent:

1. $G$ is bipartite.
2. Every cycle in $G$ has an even length.

Definition 6 (see [3]). Let $B \subseteq M_A$ be the set of moves and let $x_1, x_2 \in A^{-1}[t]$. We say that $x_2$ is accessible from $x_1$ by $B$ if there exists a sequence of moves $z_1, \ldots, z_k \in B$ such that $x_2 = x_1 + \sum_{k=1}^{K} e_k z_k$.

Remark 7 (see [4]). Let $n$ be a multiple of 3 such that $n \geq 6$, and let $x_j \in A^{-1}[t], j = 1, \ldots, k$ be the representative elements of the set of $3 \times \frac{n}{3}$-contingency tables and $B = \{ z_1, z_2, \ldots, z_k \}$ such that each $z_j$
\[ j = 1, 2, \ldots, k \] is a matrix of dimension \( 3 \times \frac{n}{3} \) either has two columns \( (1, -1,0)^t, (-1,1,0)^t, (1,0,-1)^t \) or either \( (0,1,-1)^t, (0,-1,1)^t \) and the other columns are zero denoted by \( +z_j \), or it has two columns \( (1,0,-1)^t, (1,-1,0)^t, (-1,0,1)^t, (-1,1,0)^t \) and the other columns are zero denoted by \( -z_j \). 

Also, we can write all elements of \( B \) as one-dimensional column vectors as follows:

\[
\mathbf{z}_j = (z_{i,j}, ..., z_{n,j}; j = 1, ..., k \text{ and } z_i = 1 \text{ or } -1 \text{ or } 0 \text{ such that if } t = 1, 2, ..., \frac{n}{3} \]

\[
z_t = \begin{cases} 
1 & \text{if } z_{t-3} + z_{t-2} = -1 \text{ and } \sum_{i=1}^{\frac{n}{3}} z_i = -1 \\
-1 & \text{if } z_{t-3} + z_{t-2} = 1 \text{ and } \sum_{i=1}^{\frac{n}{3}} z_i = 1 \\
0 & \text{if } z_{t-3} + z_{t-2} = 0 \text{ and } \sum_{i=1}^{\frac{n}{3}} z_i = 0 
\end{cases}
\]

Theorem 8 (see [4]). The number of elements in \( B \) equal to \( n^2 - 2n \).

Remark 9 (see [4]). Given a contingency table \( \mathbf{X} = (x_{11}, x_{12}, ..., x_{n1}, x_{1n}, ..., x_{n1}, x_{nn}) \), the entry of the matrix \( \mathbf{A} \) in the column indexed by \( x_{11}, x_{12}, ..., x_{1n} \) respectively and its rows indexed by 

\[
\sum_{i=1}^{\frac{n}{3}} x_{ij} + \sum_{j=1}^{\frac{n}{3}} x_{ij} + x_{ij} = 1, \quad x_{ij} = 0, 1, \quad j = 1, 2, ..., \frac{n}{3}
\]

respectively. The entry in the column indexed by \( x_{i,j} \) in the matrix \( \mathbf{A} \) will be equal to one if \( x_{i,j} \) appears in the index of its row, and otherwise it will be zero. Then

\[
\mathbf{A} = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 0 & \cdots & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & \cdots & 0 & 0 \\
0 & 0 & 1 & 1 & 1 & 1 & \cdots & 1 & 1 \\
0 & 1 & 0 & 0 & 0 & 1 & \cdots & 0 & 0 \\
0 & 1 & 0 & 0 & 0 & 1 & \cdots & 0 & 0 \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\
0 & 0 & 1 & 0 & 0 & 0 & \cdots & 1 & 1 \\
0 & 0 & 1 & 0 & 0 & 0 & \cdots & 1 & 1 \\
\end{bmatrix}
\]

Theorem 10 (see [4]). The set \( B = \{z_1, ..., z_{n^2 - 2n} \} \) is a set of moves.

Corollary 11 (see [4]). The set \( B \) of moves in theorem 10 is a Markov basis.

Corollary 12 (see [4]). The toric ideal \( I_A \) for \( n^2 - 2n \times 3 \times \frac{n}{3} \) - contingency tables are

\[
I_A = \langle x_{i,j} : i \neq j, j \neq k \rangle, \quad i, j = 1, 2, ..., \frac{n}{3} \text{ and } k = 0, \frac{2n}{3}, \frac{2n}{3} + \frac{n}{3}, \frac{2n}{3} + \frac{2n}{3} \text{ such that } i < j \text{ and } i < k \rangle \leq C \{P_1, P_2, ..., P_{\frac{n}{3}} \}
\]

Remark 13 (see [4]). Now, we will construct a connected graph by using the elements of \( B \). Let \( \mathbf{z}_m \) be an element of \( B \) such that \( \mathbf{z}_m = \mathbf{x}_m - \mathbf{x}_{m-1} \) and \( \mathbf{x}_m \) is an edge connected \( \mathbf{x}_m \) and \( \mathbf{x}_{m-1} \), ..., and

\[
x_{i-2} \cdot z_{i-2} + z_{i-1} = x_{0} - x_{i-2} \cdot z_{i-1} + z_{i-2} = 1
\]

we can connected all \( n^2 - 2n \times 3 \times \frac{n}{3} - \) contingency tables with fixed two dimensional marginals by \( n^2 - 2n \) edges by applying moves from \( B \) one by one and go from \( \mathbf{x}_0 \) to
without causing negative cell frequencies on the way, and also from $x_{3}^{2n-2}x_{1}$ to $x_{0}$, of this type, by forming undirected graph $G = (R, W, B) = A^{-1}[t] \cdot B$, where the contingency tables interpreted as vertices and connecting moves are interpreted as edges of a graph, $R = \{ x_{0}, x_{1}, ..., x_{3}^{2n-2} \}$ and $W = \{ x_{1}, x_{2}, ..., x_{3}^{2n-2} \}$ as shown in figure 1.

**Figure 1.** The graph $G = (R, W, B) = A^{-1}[t] \cdot B$.

In [5] H. H. Abbass and H. S. Mohammed Hussein assumed $n = 3m$, $m \in \mathbb{N}$, and $m \geq 2$, and $H$ is the subgroup $A^{-1}[t]$, where $g_{t} = (t_{1}, t_{2}, t_{3}, ..., t_{3^{2n-2}})^{t}$.

**Theorem 14** (see [4]). The graph $G = (R, W, B)$ is a connected bipartite graph (up to graph isomorphism).

**Theorem 15** (see [5]). The Markov basis $B$ is $H$-invariant.

**Corollary 16** (see [5]). The subgroup $H$ is the largest subgroup of the group $D_{3m}$ such that the Markov basis $B$ is $H$-invariant.

**Remark 17** (see [5]). Let $t = (t_{1}, t_{2}, t_{3}, ..., t_{3^{2n-2}})^{t}$.

**Theorem 18** (see [5]). If $g \in H$, then $B$ is a Markov basis for $n^{3^{2n-2}}$-contingency tables $[g]_{x_{0}, x_{1}, ..., x_{3}^{2n-2}]$ in $A^{-1}[g_{t}]$.

**Corollary 19** (see [5]). The toric ideal for $n^{3^{2n-2}}$-contingency table in $A^{-1}[g_{t}]$ is $I_{A} = \langle P_{g}^{(i+j+k)}P_{g}^{(i+j)} \rangle$, for all $g \in H$.

### 3. The Main Results

Let $n = 3m$, $m \in \mathbb{N}$ and $m \geq 2$, let $g_{x_{j}} \in A^{-1}[g_{t}]$, $j = 0, ..., n^{3^{2n-2}} - 1$, and $g \in H$ be representative elements of the set of $3 \times n^{3^{2n-2}}$-contingency tables. Then we write $g$ as $n \times n$ permutation matrix $T_{g} = [P_{ij}^{(g)}] = [\delta_{ij} \cdot g^{(j)}]$, where $\delta$ is the Kronecker's delta such that $T_{g}^{(i,j)} = T_{g}^{(i,j)} = T_{g}^{(j,)g}$, for $g \in H$, and $T_{g}^{(i,j)} = T_{g}^{(i,j)}$.

The identity matrix of the order $n$ denoted by $E_{n}$ for the unit element $e$.

Now, we consider a left action of dihedral group $D_{n}$, $n = |I|$, on $A^{-1}[t]$ the set of $n^{3^{2n-2}} \times 3 \times n^{3^{2n-2}}$ -contingency tables, and the action of dihedral group $D_{n}$ on the set of Markov basis $B$.

**Theorem 20.** Let $x_{i}, x_{j} \in A^{-1}[t]$, if $g \in H$. Then $x_{i}$ accessible from $x_{j}$ by $B$ if and only if $T_{g}x_{i}$ accessible from $T_{g}x_{j}$ by $B$, for all $i, j = 1, 2, ..., n^{3^{2n-2}}$.

**Proof:** If $x_{i}$ accessible from $x_{j}$ by $B$ then there exists a sequence of moves $x_{1}, ..., x_{k} \in B$ and $\mathbf{e}_{k} \in \{-1, 1\}$, $k = 1, ..., n^{3^{2n-2}} - 1$, such that

$$x_{i} = x_{j} + \sum_{k=1}^{n^{3^{2n-2}} - 1} e_{k}x_{k} \cdot x_{j} + \sum_{k=1}^{n^{3^{2n-2}} - 1} e_{k}x_{k} \in A^{-1}[t]$$

for $1 \leq k \leq n^{3^{2n-2}} - 1$ (Definition 6).

Let $g \in H$.

If $g = e \Rightarrow E_{n}x_{i} = E_{n}x_{j} + \sum_{k=1}^{n^{3^{2n-2}} - 1} e_{k}E_{n}x_{k}, E_{n}x_{j} + \sum_{k=1}^{n^{3^{2n-2}} - 1} e_{k}E_{n}x_{k} \in H(A^{-1}[t])$

for $1 \leq k \leq n^{3^{2n-2}} - 1$.

If $g = r^{3}$ then we write $r^{3}$ as an $n \times n$ permutation matrix, i.e. $r^{3} = T_{r}^{(1, 2, 3, n, 2, 3, 1, 3^{2n-2} + 1, 2, 3^{2n-2} + 2, 3^{2n-2} + 3, n)}$ then

$$r^{3} = T_{r}^{(1, 2, 3, n, 2, 3, 1, 3^{2n-2} + 1, 2, 3^{2n-2} + 2, 3^{2n-2} + 3, n)}$$
\[ T\left(\frac{x}{3} + 1 \frac{n-1}{3} \right)\left(\frac{n-2}{3} \frac{n-2}{3} + \frac{n}{3} \right) \cdot \left(\frac{n}{3} \right) \right] \]
If \( g = s r \frac{2n}{3} + 1 \), then
\[
\frac{2n}{3} + 1 = T_1(1) \left( 1 - \frac{2n}{3} \right) \left( \frac{2n}{3} + 1 \right) \left( \frac{2n}{3} - n - 1 \right) \left( \frac{2n}{3} + 1 \right) \left( \frac{2n}{3} - n - 1 \right)
\]
\[
T_1(1) = \frac{2n}{3} + 1 = \left( \frac{2n}{3} + 1 \right) \sum_{k=1}^{\frac{2n}{3} - 1} \epsilon_k T_1(1) \left( 1 - \frac{2n}{3} \right) \left( \frac{2n}{3} - n - 1 \right) \left( \frac{2n}{3} + 1 \right) \left( \frac{2n}{3} - n - 1 \right) \left( \frac{2n}{3} + 1 \right) \left( \frac{2n}{3} - n - 1 \right)
\]
\[
\sum_{k=1}^{\frac{2n}{3} - 1} \epsilon_k T_1(1) \left( 1 - \frac{2n}{3} \right) \left( \frac{2n}{3} - n - 1 \right) \left( \frac{2n}{3} + 1 \right) \left( \frac{2n}{3} - n - 1 \right) \left( \frac{2n}{3} + 1 \right) \left( \frac{2n}{3} - n - 1 \right)
\]
\[
T_1(1) = \frac{2n}{3} + 1 = \left( \frac{2n}{3} + 1 \right) \sum_{k=1}^{\frac{2n}{3} - 1} \epsilon_k T_1(1) \left( 1 - \frac{2n}{3} \right) \left( \frac{2n}{3} - n - 1 \right) \left( \frac{2n}{3} + 1 \right) \left( \frac{2n}{3} - n - 1 \right) \left( \frac{2n}{3} + 1 \right) \left( \frac{2n}{3} - n - 1 \right)
\]
Therefore \( T_g x_i \) accessible from \( T_g x_j \) by \( H(B) \).

Conversely, if \( T_g x_i \) accessible from \( T_g x_j \) by \( H(B) \), then there exists a sequence of moves \( g z_1, \ldots, g z_s \in H(B) \), \( g \in H \), and \( e_k \in \{-1, 1\} \), \( k = 1, \ldots, \frac{n^3 - 3n}{3} \), such that
\[
T_g x_i = T_g x_j + \sum_{k=1}^{\frac{n^3 - 3n}{3}} e_k T_g z_k,
\]
\[
T_g x_j + \sum_{k=1}^{\frac{n^3 - 3n}{3}} e_k T_g z_k \in H(A^{-1}[t]) \text{ for } 1 \leq k \leq \frac{n^3 - 3n}{3}. \quad \text{(Definition 6)}
\]
By multiplying the previous equations by \( T' \), \( g \in H \), we have
\[
T' g (T_g x_i) = T' g (T_g x_j) + \sum_{k=1}^{\frac{n^3 - 3n}{3}} e_k T' g (T_g z_k),
\]
\[
T' g (T_g x_j) + \sum_{k=1}^{\frac{n^3 - 3n}{3}} e_k T' g (T_g z_k) \in H(A^{-1}[t]) \text{ for } 1 \leq k \leq \frac{n^3 - 3n}{3}.
\]
This implies,
\[
(T' g T_g) x_i = (T' g T_g) x_j + \sum_{k=1}^{\frac{n^3 - 3n}{3}} e_k (T' g T_g) z_k,
\]
\[
(T' g T_g) x_j + \sum_{k=1}^{\frac{n^3 - 3n}{3}} e_k (T' g T_g) z_k \in H(A^{-1}[t]) \text{ for } 1 \leq k \leq \frac{n^3 - 3n}{3}.
\]
Hence \( E_n x_i = E_n x_j + \sum_{k=1}^{\frac{n^3 - 3n}{3}} e_k E_n z_k,
\]
\[
E_n x_j + \sum_{k=1}^{\frac{n^3 - 3n}{3}} e_k E_n z_k \in H(A^{-1}[t]) \text{ for } 1 \leq k \leq \frac{n^3 - 3n}{3}.
\]
Therefore, \( x_i = x_j + \sum_{k=1}^{\frac{n^3 - 3n}{3}} e_k z_k,
\]
\[
x_j + \sum_{k=1}^{\frac{n^3 - 3n}{3}} e_k z_k \in A^{-1}[t] \text{ for } 1 \leq k \leq \frac{n^3 - 3n}{3} \quad \text{[since } E_n x_j = x_j \text{ and } E_n x_k = z_k \text{]}
\]
Then \( x_i \) accessible from \( x_j \) by \( B \). \( \square \)

**Remark 21.** Now, we will construct a connected graph by using the elements of \( H(B) \). Let \( g z_k = T_g x_k \) be an element of \( H(B) \) for all \( g \in H \) such that \( g z_k = g x_k - g x_{k-1} = T_g x_k - T_g x_{k-1} \).

If \( k = 1, 2, \ldots, \frac{n^3 - 3n}{3} \) is an edge connected \( g z_k = T_g x_k \) and \( g z_{k-1} = T_g x_{k-1} \) and, \( \ldots \), and \( g z_{\frac{n^3 - 3n}{3}} = g x_0 - g x_{\frac{n^3 - 3n}{3} - 1} = T_g x_0 - T_g x_{\frac{n^3 - 3n}{3} - 1} \) be an...
edge connected \( T_g x_0 \) and \( T_g x_{2-3n} \),
where \( g x_i \in A^{-1}[g] \). Then we can connect all \( n^{3-2n} \times 3 \) contingency tables with fixed two dimensional marginals by \( n^{3-2n} \) edges by applying moves from \( H(B) \) to \( T_g x_0 \) one by one and go from \( T_g x_2 \) to \( T_g x_{2-3n} \) without causing negative cell frequencies on the way, and also from \( T_g x_{2-3n} \) to \( T_g x_0 \).

This forms undirected graph as shown in figure 2

![Figure 2](image-url)

Figure 2. The graph \( G = (T_g R, T_g W, H(B)) \), where the contingency tables interpreted as vertices and connecting moves are interpreted as edges of a graph.

**Theorem 22.** The graphs \( T_g G = (T_g R, T_g W, H(B)) \) are connected bipartite graphs (up to graph isomorphism).

**Proof:**
Let \( T_g x_i, T_g x_j \in A^{-1}[g] \), if
\[
0 \leq i \leq j \leq \frac{n^{3-2n}}{3} - 1, i \neq j,
\]
then remark 21 there exists a path \( < T_g x_i, T_g x_j, T_g x_{j+1}, T_g x_{j+2}, \ldots, T_g x_{j-1}, T_g x_j > \), and if \( 0 \leq j \leq i \leq \frac{n^{3-2n}}{3} - 1, i \neq j \), by remark 21 there exists a path \( < T_g x_j, T_g x_{j+1}, T_g x_{j+2}, \ldots, T_g x_{i-1}, T_g x_i > \), and that implies there exists a path between every pair of distinct vertices \( T_g x_i, T_g x_j \in A^{-1}[g] \) of the graph, by (definition 3), \( G \) is a connected graph.

Now, we prove the graph \( G = (T_g R, T_g W, H(B)) \) is a bipartite graph.

Let \( T_g x_i, T_g x_j, T_g x_{i-1}, T_g x_{j-1}, T_g x_{i+j}, T_g x_{j+i} \in T_g X \) be a cycle in \( G \). Suppose \( T_g x_i \in T_g R \) and then \( T_g x_{i+1} \in T_g W \) since the edge \( T_g x_{i+1} = T_g x_{i+1} - T_g x_i \in H(B) \), then \( T_g x_{i+2} \in T_g R \), and so on.

We see that if \( k \) is an odd, then \( T_g x_k \in T_g W \), and if \( k \) is even, then \( T_g x_k \in T_g R \). Since \( T_g x_{j+1} = T_g x_{j+1} - T_g x_j \) it implies that \( j + 1 \) is even and thus the cycle is of even length. By theorem 5, then the graph \( G = (T_g R, T_g W, H(B)) \) is a bipartite graph.

**4. Genomics and Phylogenetic**

In this section, we describe some of the basic biological facts needed to understand phylogenetic models and then delve into the practical side of the algebraic statistics of these models. The basic genetic information of an organism is (almost always) carried in the form of DNA, a double helix consisting of two complementary B polymers bound together. The DNA molecules in a genome are typically represented as a number of frequencies of letters from the four letters alphabet \( \{ A, C, G, T \} \). These letters correspond to the bases in the double helix that is the nucleotides Adenine, Cytosine, guanine and Thymine. The four nucleotides that form DNA come in two types: the purines (A and G) and the pyrimidine's (C and T). The two strands of the double helix are joined together via the base pairings A to T (via 2 hydrogen bonds) and C to G (via 3 hydrogen bonds). Since each cell typically contains a copy of the DNA of the organism, DNA copying occurs frequently. Several types of errors are possible during the replication of DNA. Single bases can mutate, or large pieces of DNA can separate and become reattached, possibly at another position, possibly in the opposite direction, these are just some of the events that occur over the course of evolution (C. Semple, M [2] and J. Felsenstein [6]).

**5. A New Model of Genetic Algorithm Using the Action of Largest Subgroup of \( D_n \) for Invariance Markov Basis**

In this section, we construct a new model of genetic algorithm that permutes the pieces of nucleotides in aligned DNA sequences using the actions of largest subgroup \( H \) of \( D_n \) for invariance Markov basis and toric ideals. Now, we describe our model in the following steps.

**Step (1):** Suppose we have \( t \)-taxons of DNA sequences each taxon of length \( L \) such as

**Taxon1:** A G C T A A C G T A T

**Taxon2:** C G A T C T G A C C T T

**Taxon3:** A C G T C A C G T A G C

Now, we define a pattern \( \ell = i_1, i_2, \ldots, i_m \) to be the sequence of characters. We look at a single site (column) of our sequence data. In the sequences above, we can look at the first site in the sequences and see the pattern "AC . . . A". A pattern frequency \( x_{j1} \) is that \( x \) appears in our set of sequence data, and we denote to the number of frequencies by \( x \), where \( n = 3^m, m \in \mathbb{N}, m \geq 2 \).

**Step (2):** We can input pattern frequency \( x_{i1} \) of above sequences in \( n^3 \times \frac{n}{3} \)-contingency table as follows:
Where \( x = (x_1, x_2, \ldots, x_n) \) is a vector of non-negative integers and \( \mathbf{A} = (a_{ij}) \) is a matrix, then the \( \mathbf{A} \)-fiber is defined as \( \mathbf{A}^{-1} \{ t \} = \{ x \in \mathbb{N}^n : \mathbf{Ax} = t \} \).

Step (3): Represent the contingency table as a \( \mathbf{A} \)-dimensional column vector of non-negative integers.

Step (4): From remark 9, \( \mathbf{A} \) is an \( \frac{n+q}{2} \times n \) matrix and \( \mathbf{A}^\top \) is the transpose of a vector or matrix, as in remark (1.4.2) then \( \mathbf{A} \) is an \( \mathbf{A} \)-fiber (i.e., \( \mathbf{x} \in \mathbb{N}^n \)).

Step (5): We can find the Markov basis from remark 7.

Step (6): \( g = g_{1/n} = g_{1/3} \) is a linear transformation, \( t \in \mathbb{Z}^{2n} \), and \( g^{-1} \{ t \} \) be the set of \( g \)-fibers, and \( B = \ker g(A) \). Use remark 13 to find the bipartite graph \( B = \{ R, W, B \} = A^{-1} \{ t \} \).

Step (7): We can find the toric ideals by using corollary 12 for each contingency tables.

Step (8): Find \( g \in H \) where \( g = g_{1/n} = g_{1/3} \) is a permutation matrix of \( g \).

Step (9): Use corollary 21 and theorem 22 to find the graphs.

Step (10): Use corollary 19 to find the toric ideal \( I_g = \mathbb{C}[g] \).

Step (11): Use the \( \frac{n^2-2n}{3} \times 3 - \)contingency tables \( (t^2, t^3) \) in step (6), and \( \frac{n^2-2n}{3} \times 3 - \)contingency tables \( (g t) \) in step (9) for all \( g \in H \) to find the permutation of nucleotides in aligned DNA sequences.

Example (23): Suppose we have the following three aligned DNA sequences...
Taxon 1: A G C T G A T T G G C C G T T T T
Taxon 2: A G A T C T A T C G A A A C A A A T T C T C
Taxon 3: A T T C A G A T C T T T T T A A A A T T C

Step (1): There are three taxons of above DNA sequences with $|\mathcal{X}| = \sum_{i=1}^{3} x_i = 12 = 20$ and nine patterns A A A, G G T, C A T, T T C, G C A, A T G, T A A, T T T, G C C with frequencies $1, 2, 4, 1, 2, 1, 4, 3$ and $2$ respectively where $n = 9$.

Step (2): Now, we input the patterns frequency $x_i$ of above sequences in $3 \times 3$ contingency table as follows:

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>7</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>2</td>
<td>9</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>2</td>
<td>9</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>6</td>
<td>7</td>
<td>20</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Then the table of marginal and conditional probability is:

<table>
<thead>
<tr>
<th></th>
<th>0.05</th>
<th>0.10</th>
<th>0.20</th>
<th>0.35</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.05</td>
<td>0.10</td>
<td>0.20</td>
<td>0.35</td>
</tr>
<tr>
<td>2</td>
<td>0.05</td>
<td>0.10</td>
<td>0.05</td>
<td>0.20</td>
</tr>
<tr>
<td>3</td>
<td>0.20</td>
<td>0.15</td>
<td>0.10</td>
<td>0.45</td>
</tr>
</tbody>
</table>

Step (3): Represent the contingency table $\mathbf{x} = \{x_i\}_{i=1}^{3} \in \mathbb{N}^9$ as a 9-dimensional column vector of non-negative integers $\mathbf{x} = (1, 2, 4, 1, 2, 1, 4, 3, 2)$, as in remark (1.5.2) then $\mathbf{x}$ is a $t$-fiber (i.e. $\mathbf{x} \in A^{-1}[t]$, where

$$
A^{-1}[t] = \{x \in \mathbb{N}^9 : A \mathbf{x} = t\}
$$

Then the number of moves is $n^3 - 2n^2 = \frac{n^3 - 2n^2}{2} = 18$ elements in the set $A^{-1}[t]$. We can find the Markov basis from remark 7. Then $z_1 = (1, -1, 0, -1, 1, 0, 0, 0, 0)$, $z_2 = (0, 0, 0, 1, -1, 0, 1, 0, 0)$, $z_3 = (1, 0, -1, -1, 0, 1, 0, 0, 0)$.
Step (6): The connected graph $A^{-1}f_B = G = (R, W, B)$ with $\frac{n^2-3n}{2} = 18$ (t- fibres) $3 \times 3$ contingency tables as vertices of it.

Figure 3: The graph $G = (R, W, B) = A^{-1}f_B$ where the contingency tables interpreted as vertices and connecting Markov basis are interpreted as the edges of a graph $R = \{x_0, x_1, x_3, x_5, x_7, x_9, x_{11}, x_{13}, x_{15}, x_{17}\}$ and $W = \{x_1, x_2, x_4, x_6, x_8, x_{10}, x_{12}, x_{14}, x_{16}\}$.

Where
Step (7): We can find the toric ideals by using corollary 12 for each contingency table. Then the toric ideals that correspond the Markov basis that shown in the a previous figure is:

\[
I_A = \langle P_1 P_2 - P_2 P_3, P_3 P_0 - P_0 P_2, P_2 P_2 - P_2 P_4, P_0 P_0 - P_0 P_2, P_2 P_2 - P_2 P_3, P_0 P_2 - P_2 P_3, P_2 P_2 - P_0 P_2, P_2 P_2 - P_1 P_2, P_2 P_2 - P_0 P_2, P_1 P_2 - P_0 P_2 \rangle.
\]

\[
I_A = \langle 0.005 - 0.005, 0.0075 - 0.02, 0.0025 - 0.01, 0.005 - 0.01, 0.005 - 0.02, 0.01 - 0.03, 0.01 - 0.0075, 0.0075 - 0.02, 0.005 - 0.04 \rangle.
\]

\[
\Rightarrow I_A = \langle 0.0025 \rangle.
\]

Step (8): We find the set

\[T_r \cap B = \{T_r \cap z_1, T_r \cap z_2, T_r \cap z_3, T_r \cap z_4, T_r \cap z_5, T_r \cap z_6, T_r \cap z_7, T_r \cap z_8, T_r \cap z_9, T_r \cap z_{10}, T_r \cap z_{11}, T_r \cap z_{12}, T_r \cap z_{13}, T_r \cap z_{14}, T_r \cap z_{15}, T_r \cap z_{16}, T_r \cap z_{17}, T_r \cap z_{18}\}.\]

Such that
Step (9): We find the set $\{T_{r^*}x_0, T_{r^*}x_1, T_{r^*}x_2, T_{r^*}x_3, T_{r^*}x_4, T_{r^*}x_5, T_{r^*}x_6, T_{r^*}x_7, T_{r^*}x_8, T_{r^*}x_9, T_{r^*}x_{10}, T_{r^*}x_{11}, T_{r^*}x_{12}, T_{r^*}x_{13}, T_{r^*}x_{14}, T_{r^*}x_{15}, T_{r^*}x_{16}, T_{r^*}x_{17}\} \subseteq \Lambda^{-1}[r^2d] (r^2d$-fibers) where $r^2d = (9, 7, 4, 6, 7, 7)$. We have:

\[
T_{r^*}x_0 = \begin{bmatrix} 4 & 3 & 2 & 9 \\ 1 & 2 & 4 & 7 \\ 6 & 7 & 7 & 20 \end{bmatrix}, \quad T_{r^*}x_1 = \begin{bmatrix} 4 & 3 & 2 & 9 \\ 0 & 3 & 1 & 4 \\ 5 & 7 & 7 & 20 \end{bmatrix}, \quad T_{r^*}x_2 = \begin{bmatrix} 3 & 4 & 2 & 9 \\ 2 & 1 & 4 & 7 \\ 6 & 7 & 7 & 20 \end{bmatrix},
\]

\[
T_{r^*}x_3 = \begin{bmatrix} 3 & 4 & 2 & 9 \\ 1 & 2 & 4 & 7 \\ 6 & 7 & 7 & 20 \end{bmatrix}, \quad T_{r^*}x_4 = \begin{bmatrix} 3 & 4 & 2 & 9 \\ 0 & 3 & 1 & 4 \\ 5 & 7 & 7 & 20 \end{bmatrix}, \quad T_{r^*}x_5 = \begin{bmatrix} 2 & 4 & 3 & 9 \\ 3 & 1 & 5 & 7 \\ 6 & 7 & 7 & 20 \end{bmatrix},
\]

\[
T_{r^*}x_6 = \begin{bmatrix} 2 & 4 & 3 & 9 \\ 1 & 2 & 4 & 7 \\ 6 & 7 & 7 & 20 \end{bmatrix}, \quad T_{r^*}x_7 = \begin{bmatrix} 1 & 3 & 5 & 9 \\ 3 & 3 & 1 & 7 \\ 6 & 7 & 7 & 20 \end{bmatrix}, \quad T_{r^*}x_8 = \begin{bmatrix} 1 & 3 & 5 & 9 \\ 1 & 2 & 4 & 7 \\ 6 & 7 & 7 & 20 \end{bmatrix},
\]

Step (10): Use remark 21 and theorem 22 to find the graph

\[
T_{r^*}x_9 = \begin{bmatrix} 0 & 3 & 6 & 9 \\ 5 & 2 & 0 & 7 \\ 6 & 7 & 7 & 20 \end{bmatrix}, \quad T_{r^*}x_{10} = \begin{bmatrix} 0 & 3 & 6 & 9 \\ 4 & 2 & 0 & 7 \\ 6 & 7 & 7 & 20 \end{bmatrix}, \quad T_{r^*}x_{11} = \begin{bmatrix} 1 & 2 & 6 & 9 \\ 4 & 2 & 0 & 7 \\ 6 & 7 & 7 & 20 \end{bmatrix},
\]

\[
T_{r^*}x_{12} = \begin{bmatrix} 1 & 2 & 6 & 9 \\ 5 & 3 & 1 & 7 \\ 6 & 7 & 7 & 20 \end{bmatrix}, \quad T_{r^*}x_{13} = \begin{bmatrix} 2 & 2 & 5 & 9 \\ 3 & 3 & 1 & 7 \\ 6 & 7 & 7 & 20 \end{bmatrix}, \quad T_{r^*}x_{14} = \begin{bmatrix} 2 & 2 & 5 & 9 \\ 3 & 3 & 1 & 7 \\ 6 & 7 & 7 & 20 \end{bmatrix},
\]

\[
T_{r^*}x_{15} = \begin{bmatrix} 2 & 2 & 5 & 9 \\ 3 & 3 & 1 & 7 \\ 6 & 7 & 7 & 20 \end{bmatrix}, \quad T_{r^*}x_{16} = \begin{bmatrix} 2 & 4 & 3 & 9 \\ 3 & 1 & 3 & 7 \\ 6 & 7 & 7 & 20 \end{bmatrix}, \quad T_{r^*}x_{17} = \begin{bmatrix} 1 & 3 & 2 & 4 \\ 1 & 2 & 3 & 1 \\ 6 & 7 & 7 & 20 \end{bmatrix}.
\]
Figure 4: The graph \( T_{3,3} G = (T_{3,3} R, T_{3,3} W, H(B)) = A^{-1}[r^2 t]_{L(B)} = A^{-1}[r^2 t]_B \) where the contingency tables interpreted as vertices and connecting moves are interpreted as edges of a graph, and
\[
T_{3,3} R = \{ r^2 x_0, r^2 x_2, r^2 x_4, r^2 x_6, r^2 x_8, r^2 x_{10}, r^2 x_{12}, r^2 x_{14}, r^2 x_{16} \}
\]
and
\[
T_{3,3} W = \{ r^2 x_1, r^2 x_3, r^2 x_5, r^2 x_7, r^2 x_9, r^2 x_{11}, r^2 x_{13}, r^2 x_{15}, r^2 x_{17} \}.
\]

Step (11):
Use corollary 19 to find the toric ideal \( I_A = \langle p_i + r_j + k \rangle \), \( i = 1, \ldots, n \) and \( l, k = 0, \frac{2m}{3}, \frac{2m}{3} \), such that
\[
\forall i < j \text{ and } l < k \Rightarrow \langle P_i P_j \rangle \neq \langle P_l P_k \rangle.
\]

Also, we find the set \( \{ T_{s} x_5, T_{s} x_6, T_{s} x_7, T_{s} x_8, T_{s} x_9, T_{s} x_{10}, T_{s} x_{11}, T_{s} x_{12}, T_{s} x_{13}, T_{s} x_{14}, T_{s} x_{15}, T_{s} x_{16}, T_{s} x_{17} \} \) \( \subseteq A^{-1}[srt](srt-fibers) \) where, \( srt = (9, 4, 7, 7, 7, 6) \), such that
Now, use remark 21 and theorem 22 to find the graph $T_{sr} G = A^{-1}[srt]_B$.

**Step (12):** Use the $18 \times 3 \times 3$-contingency tables ($t$-fibers) in step (6), $18 \times 3 \times 3$-contingency tables ($gt$-fibers) in step (9) to find the permutation of nucleotides in aligned DNA sequences. Then the change in the type of DNA sequences under the Markov basis.

Be as Figure 3 where...
And the change in the type of DNA sequences under the action of $\mathbf{p}^3$ on the set of Markov basis. Be as Figure 4 where
And the change in the type of DNA sequences under the action of $s_{IR}$ on the set of Markov basis. Be as Figure 5 where

$$T_{s_{IR}}x_8 = \begin{bmatrix}
1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\
T_{s_{IR}}x_1 = \begin{bmatrix}
1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\
\end{bmatrix}
$$

$$T_{s_{IR}}x_9 = \begin{bmatrix}
1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\
T_{s_{IR}}x_2 = \begin{bmatrix}
1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\
\end{bmatrix}
$$

$$T_{s_{IR}}x_{10} = \begin{bmatrix}
1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\
T_{s_{IR}}x_3 = \begin{bmatrix}
1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\
\end{bmatrix}
$$

$$T_{s_{IR}}x_{11} = \begin{bmatrix}
1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\
T_{s_{IR}}x_4 = \begin{bmatrix}
1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\
\end{bmatrix}
$$

$$T_{s_{IR}}x_{12} = \begin{bmatrix}
1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\
T_{s_{IR}}x_5 = \begin{bmatrix}
1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\
\end{bmatrix}
$$

$$T_{s_{IR}}x_{13} = \begin{bmatrix}
1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\
T_{s_{IR}}x_6 = \begin{bmatrix}
1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\
\end{bmatrix}
$$

$$T_{s_{IR}}x_{14} = \begin{bmatrix}
1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\
T_{s_{IR}}x_7 = \begin{bmatrix}
1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\
\end{bmatrix}
$$

$$T_{s_{IR}}x_{15} = \begin{bmatrix}
1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\
T_{s_{IR}}x_8 = \begin{bmatrix}
1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\
\end{bmatrix}
$$

$$T_{s_{IR}}x_{16} = \begin{bmatrix}
1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\
T_{s_{IR}}x_9 = \begin{bmatrix}
1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\
\end{bmatrix}
$$

$$T_{s_{IR}}x_{17} = \begin{bmatrix}
1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\
T_{s_{IR}}x_{10} = \begin{bmatrix}
1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\
\end{bmatrix}
$$

$$T_{s_{IR}}x_{18} = \begin{bmatrix}
1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\
T_{s_{IR}}x_{11} = \begin{bmatrix}
1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\
\end{bmatrix}
$$

$$T_{s_{IR}}x_{12} = \begin{bmatrix}
1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\
T_{s_{IR}}x_{12} = \begin{bmatrix}
1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\
\end{bmatrix}
$$

$$T_{s_{IR}}x_0 = \begin{bmatrix}
1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\
\end{bmatrix}
$$
Remark 24:

i. We refer to \( ①, ②, ③, ④, ⑤ \) and \( ⑥ \) in example 23 to the frequencies of the patterns in DNA sequences.

ii. We refer \( ⓿ \) to the hidden in the pattern frequency of DNA sequences.

References


