Toric Ideals for $2 \binom{n}{\frac{n}{2}} \times 2 \times \frac{n}{2}$ - Contingency Tables with Fixed Two Dimensional Marginals When $n$ is an Even Number

Husein Hadi Abass¹, Yahya .Mahdy.AL-Mayaly², and Bashar Ahmed Jawad Sharba³

Faculty of Mathematics and Computer Science
University of Kufa
Najaf - Kufa – Iraq

Abstract:

In this paper, we find Markov basis and toric ideals for $2 \binom{n}{\frac{n}{2}} \times 2 \times \frac{n}{2}$ - contingency tables with fixed two dimensional marginals when $n$ is an even number greater than or equal to 4 and we give an application of $2 \binom{n}{\frac{n}{2}} \times 2 \times \frac{n}{2}$ -contingency tables in genetic algorithm.

Key words and phrases: Computational algebraic statistics, Sufficient statistics, Linear transformation, Connected graph, Bipartite graph.

1. Introduction:

Since In 1998 the publication of P.Diaconis and B.Sturmfels, the new field of computational algebraic statistics has been developing rapidly, and in the same year P.Diaconis and B.Sturmfels defined the notion of Markov basis for constructing a connected Markov chain for sampling from a conditional distribution over a discrete sample space and proved the fundamental fact that a Markov basis corresponds to a set of binomial generators of a toric ideal[9]. In 2000, M. Dyer, and C. Greenhill, found a Polynomial-time counting and sampling of two-rowed contingency tables[8]. In 2001, A.Dobra showed that the only moves that have to be included in a Markov basis that links all contingency tables having a set of fixed marginals when this set of marginals induces a decomposable graphical models[1]. In 2002, A. Dobra, and S.Sullivant, described a divide-and-conquer algorithm for generating Markov basis of multi-way tables that connects all tables of counts having a fixed set of marginal totals[2]. In 2003, S. Aoki, and A.Takemura, proved...
that there exist a unique minimal basis for $3 \times 3 \times K$ contingency tables consisting of four types of indispensable moves [13]. And in the same year S. Aoki, and A. Takemura presented a list of indispensable moves of unique minimal Markov basis of $3 \times 4 \times K$ and $4 \times 4 \times K$ contingency tables with fixed two-dimensional marginals[12], also A. Takemura, and S. Aoki given some characterizations of minimal Markov basis for connected Markov chain and given a necessary and sufficient condition for uniqueness of minimal Markov basis [3]. In 2005, A. Takemura, and S. Aoki. Studied the Markov basis for sampling from discret sample space, which is equipped with some convet metric and they started from two state in the sample space, and they asked whether they can always move closer by an element of a Markov basis and they called a Markov basis distance reducing[4]. In this paper, we give a new algorithm to find the Markov basis and toric ideals for $(\ )$-contingency tables with fixed two dimensional marginals.

**2. Some Basic Concepts:**

In this section, we review some basic definitions and notations of contingency tables, moves, Markov basis, and toric ideals that we need in our work.

**Definition (2.1):** [11]

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 $x_i \in \mathbb{N} = \{1, 2, \ldots \}$ denotes the frequency of a cell $i$. The set of frequencies is called a **contingency table** and denoted as $x = \{x_i\}_{i \in I} \in \mathbb{N}^n$ 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 $i \mapsto x_i$.

**Remark(2.2):** [11]

The $L_1$-norm of $x \in \mathbb{N}^n$ is called the **sample size** and denoted as $|x| = \sum_{i \in I} x_i$. We will denote $\mathbb{Z}$ to the set of integer numbers, also we denote to the set of integer numbers, also we denote to the $a_j \in \mathbb{Z}^n$, $j = 1, \ldots, v$, as fixed column vectors consisting of integers. A $v$-dimensional column vector $t = (t_1, \ldots, t_v)' \in \mathbb{Z}^v$ as $t_j = a_j' x$, $j = 1, \ldots, v$. Here $'$ 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 

$$A = \begin{bmatrix} a'_{j1} \\
\vdots \\
a'_{jp} \end{bmatrix},$$

and if $t = Ax$ is a $v$-dimensional column vector, we define the set $\theta = \{t: t = Ax, x \in \mathbb{N}^p\} = AN^p \subset \mathbb{Z}^v$, where $\mathbb{N}$ is the set of natural numbers. In typical situations of a statistical theory, $t$ is sufficient statistic for the nuisance parameter. The set of $x$'s for a given $t$, $A^{-1}[t] = \{x \in \mathbb{N}^p: Ax = t\}$ ($t$-fibers), is considered for performing similar tests, for the case of 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 $x$'s with the same row sums and column sums to $t$. The set of $t$-fibers gives a decomposition of $\mathbb{N}^p$. 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 $x_1 \sim x_2 \iff x_1 - x_2 \in \ker(A)$. This relation is an equivalence relation and $\mathbb{N}^p$ 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 a label for these equivalence classes.

**Definition (2.3):**

An $n$-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., $Az = 0$.

**Remark (2.4):**

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 $z^+, z^- \in \mathbb{N}^n$. 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 $M = M_A = \mathbb{Z}^n \cap \ker(A)$. 

118
Definition (2.5): [3]

Let $B \subset M_n$ be the set of moves and let $x_1, x_2 \in A^{-1}[t]$. We say that $x_2$ accessible from $x_1$ by $B$ if there exists a sequence of moves $z_1, \ldots, z_s \in B$ and $\varepsilon_s \in \{-1, 1\}$, $s = 1, \ldots, S$, such that

$$x_2 = x_1 + \sum_{s=1}^{S} \varepsilon_s z_s,$$

$$x_1 + \sum_{s=1}^{S} \varepsilon_s z_s \in A^{-1}[t] \text{ for } 1 \leq r \leq S.$$

Now, we give some concepts about graph theory that we use later

Definition (2.6): [6]

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 (2.7): [6]

A graph $G$ is bipartite graph if there are $X, Y \subseteq V(G)$ meeting the following conditions:

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

Theorem (2.8): [6]

For a graph $G$ the following statements are equivalent:

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

Definition (2.9): [10]

Let $A: \mathbb{Z}^n \to \mathbb{Z}^v$ a linear transformation, $t \in \mathbb{Z}^v$, and $A^{-1}[t]$ be the set of $t$-fibers, and let $B \subset \ker_A$ , then we define $A^{-1}[t]_B$ be the graph with vertex set $A^{-1}[t]$ and $u - v$ an edge if and only if $u - v \in \pm B$.

Definition (2.10):[3]

If $B \subset \ker_A$ is a set such that $A^{-1}[t]_B$ is connected for all $t$, then $B$ is a Markov basis for $A$.

Remark (2.11):

Throughout this paper the symbol $K$ denotes a field of complex numbers $\mathbb{C}$,

The set $K^P$ is the vector space of tuples of elements in $K$.

Henceforth $P_1, P_2, \ldots, P_n$ denote indeterminates, that is, polynomial variables.

A monomial $m$ in the indeterminates $P_1, P_2, \ldots, P_n$ is a expression of the form $m = \prod_{i=1}^{p} P^{a_i}$,
where $\alpha_1, \alpha_2, \ldots, \alpha_n$ are nonnegative integers. We will often use the shorthand

$$m = \prod_{i=1}^{p} P_{\alpha_i} = P^\alpha$$

to denote this monomial. A polynomial is a linear combination of finitely many $K^P$ monomials

$$f(p) = f(P_1, P_2, \ldots, P_n) = \sum_{\alpha} c_\alpha P^\alpha$$

where the $c_\alpha \in K$ and at most finitely many of them are nonzero. Note any polynomial $f(P)$ is also a function from $K^p$ to $K$, simply by evaluating the polynomial at a point of $K^p$.

The set of all polynomials in the $p$ indeterminates $P_1, P_2, \ldots, P_p$ is denoted by either

$$K[P_1, P_2, \ldots, P_p]$$

for short. Note that $K[p]$ has the structure of a ring because we can add and multiply two polynomials to produce new polynomials, and these addition and multiplication operations are well-behaved with respect to one another.

**Definition (2.12):** [10]

Let $A: \mathbb{Z}^n \rightarrow \mathbb{Z}^d$ be a linear transformation, the Toric ideal $I_A$ is the ideal $< P^u - P^v : u, v \in \mathbb{N}^n, A(u) = A(v) > \subseteq K[P_1, \ldots, P_p]$ where $P^u = P_1^{u_1} P_2^{u_2} \ldots P_p^{u_p}$.

**Theorem (2.13) (Diaconis - Sturmfels 1998):** [3]

A collection of binomials $\{p^{z^+} - p^{z^-} : z \in B\} \subseteq I_A$ is a generating set of the toric ideal $I_A$ if and only if $\pm B$ is a Markov basis for $A$.

In particular, since every toric ideal has a finite generating set of binomials, Markov basis exist.

### 3. Genomics and phylogenetics

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 four nucleotides that form DNA come in two types: the purines (A and G) and the pyrimidines (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.[5],[7].

4. The Main Results
In this section, we find Markov basis and toric ideal for $2 \binom{n}{2} \times 2 \times \frac{n}{2}$ - contingency tables with fixed two dimensional marginals when $n$ is an even number greater than or equal to 4. Also, we construct a new model of genetic algorithm that permutes the pieces of Nucleotides in aligned DNA sequences by using Markov basis and toric ideals for a previous contingency tables.

Remark(4.1):
Let $n$ be an even number such that $n \geq 4$, and let $x_j \in A^{-1}[i]$, $j = 1, ..., k$ be the representative elements of the set of $2 \times \frac{n}{2}$ - contingency tables and $B = \{z_1, z_2, ..., z_k\}$ such that each $z_m, m = 1, 2, ..., k$, is a matrix of dimension $2 \times \frac{n}{2}$ either has two columns $(1, -1)^\prime, (-1, 1)^\prime$ and the other columns are zero denoted by $+z_m$, or it has two columns $(-1, 1)^\prime, (1, -1)^\prime$ and the other columns are zero denoted by $-z_m$ like

$$\begin{pmatrix} 1 & -1 & 0 \\ -1 & 1 & 0 \end{pmatrix}, \begin{pmatrix} -1 & 1 & 0 \\ 1 & -1 & 0 \end{pmatrix}.$$ 

Also, we can write all elements of $B$ as one dimensional column vector as follows:

$$z_m = (z_1, ..., z_n)^\prime, m = 1, ..., k \text{ and } z_s = 0 \text{ or } 1, s = 1, 2, ..., n \text{ such that}$$

$$z_s =$$

$$\begin{cases} 1 \text{ if } x_{i+s} = -1 \text{ and } \sum_{i=1}^{n-s} z_i = -1 \text{ and } \sum_{i=s+1}^{n} z_i = 1 \\ -1 \text{ if } x_{i+s} = 1 \text{ and } \sum_{i=1}^{n-s} z_i = 1 \text{ and } \sum_{i=s+1}^{n} z_i = -1 \\ 0 \text{ if } x_{i+s} = 0 \text{ and } \sum_{i=1}^{n-s} z_i = \sum_{i=s+1}^{n} z_i = 0 \end{cases}$$

$(2.1)$.

Theorem (4.2):
The number of elements in $B$ equal to $2 \binom{n}{2}$.

Proof:
From remark(4.1), there are two rows and $\frac{n}{2}$ columns in $+z_m$, such that it has two columns $(1, -1)^\prime, (-1, 1)^\prime$ and the other
columns are zero, then the number of elements \( +z_m \) in \( B \) is \( \left( \frac{n}{2} \right) \), but \(-z_m\) is an element in \( B \) for all \( m = 1, 2, ..., k \). Then the number of elements \(-z_m\) in \( B \) is \( \left( \frac{n}{2} \right) \). Since each element in \( B \) is either \(+z_m\) or \(-z_m\), then the numbers of elements in \( B \) is \( \left( \frac{n}{2} \right) + \left( \frac{n}{2} \right) = 2 \left( \frac{n}{2} \right) \).

Now, we will prove all elements in \( B \) are moves.

**Remark (4.3):**

**Theorem (4.4):**

The set \( B = \{ z_1, ..., z_m \} \) is a set of moves.

**Proof:** Let \( z_m \in B \). To prove \( z_m \) is a move.

By remark (4.3)

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

We must show that

Given a contingency table \( x \), the entry of the matrix \( A \) in the column indexed by \( x = (x_1, x_2, ..., x_n) \) and row \( \left( \sum_{i=1}^{n} x_i, \sum_{i=2}^{n+1} x_i, x_1 + x_{n+2}, ..., x_2 + x_n \right) \) will be equal to one if \( x_i \) a pears in the \( (\sum_{i=1}^{n} x_i) \) and it will zero otherwise. Then

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

From (2.1), we have

if \( i = 1 \) \( \Rightarrow \) \( \sum_{j=1}^{n} a_{1j} z_j = \sum_{j=1}^{n} z_j = 0 \).

if \( i = 2 \) \( \Rightarrow \) \( \sum_{j=1}^{n} a_{2j} z_j = \sum_{j=2}^{n+1} z_j = 0 \).

if \( i = 3, ..., A + n \) \( \Rightarrow \) \( \sum_{j=1}^{n} a_{ij} z_j = z_j + z_{n+j} = 0 \), \( \forall i, \forall j \).
Then

\[ A_{x_m} = \begin{bmatrix}
1 & 1 & 0 & 0 & \cdots & 0 \\
0 & 0 & 1 & 1 & \cdots & 1 \\
1 & 0 & 0 & 0 & \cdots & 1 \\
0 & 1 & 0 & 0 & \cdots & 1 \\
0 & 0 & 1 & 0 & \cdots & 1 \\
0 & 0 & 0 & 1 & \cdots & 1 \\
0 & 0 & 0 & 0 & \cdots & 1 \\
\end{bmatrix},
\]

Therefore, \( B \subseteq \ker A \). This implies that \( B \) is a set of moves (By definition (2.3)).

**Remark (4.5):**

Now, we will construct a connected graph by using the elements of \( B \). Let \( z_m \) be an element of \( B \) such that \( z_m = x_m - x_{m-1} \), \( m = 1, 2, \ldots, 2 \left( \frac{n}{2} \right) \) is an edge connected \( x_m \) and \( x_{m-1} \), \( \ldots \), and \( z_{2\left( \frac{n}{2} \right)} = x_0 - x_{2\left( \frac{n}{2} \right)-1} \) is an edge connect \( x_0 \) and \( x_{2\left( \frac{n}{2} \right)-1} \), where \( x_t \in A^{-1}[t], i = 0, 2, \ldots, 2 \left( \frac{n}{2} \right) - 1 \). Then we can connected all \( 2 \left( \frac{n}{2} \right) \times 2 \times \frac{n}{2} \) contingency tables with fixed two dimensional marginals by \( 2 \left( \frac{n}{2} \right) \) edges by applying moves from \( B \) to \( x_0 \) one by one and go from \( x_0 \) to \( x_{2\left( \frac{n}{2} \right)-1} \) without causing negative cell frequencies on the way, and also from \( x_{2\left( \frac{n}{2} \right)-1} \) to \( x_0 \), of this type, by forming undirected graph as shown in figure (1).

![Graph](image)

**Figure (1): The graph \( G = (R, W, B) = A^{-1}[t]_B \), Where the contingency...**
tables interpreted as vertices and connecting moves are interpreted as edges of a graph, \( R = \{ x_0, x_2, ..., x_{\binom{n}{2}-2} \} \) and \( W = \{ x_1, x_3, ..., x_{\binom{n}{2}-1} \} \).

**Theorem(4.6):**

The graph \( G = (R, W, B) \) is a connected bipartite graph (up to graph isomorphism).

**Proof:**

To prove \( G = (R, W, B) \) is a connected graph.

Let \( x_i, x_j \in A^{-1}[t] \),

if \( 0 \leq i, j \leq \left( \frac{n}{2} \right) - 1 \) or \( \left( \frac{n}{2} \right) - 1 < i, j \leq 2 \left( \frac{n}{2} \right) - 1, i \neq j \), there exists a path
\[
<x_i, z_{i+1}, x_{i+1}, z_{i+2}, ..., x_{j-1}, z_j, x_j>
\]
and if \( 0 \leq i, j \leq \left( \frac{n}{2} \right) - 1 \) and \( \left( \frac{n}{2} \right) - 1 < i, j \leq 2 \left( \frac{n}{2} \right) - 1, i \neq j \)
there exists a path
\[
x_i, z_{i+1}, x_{i+1}, z_{i+2}, ..., x_{j-1}, z_j, x_j
\]
and that implies there exists a path between every pair of distinct vertices \( x_i, x_j \in A^{-1}[t] \) of the graph \( G \), by (definition (2.6)), \( G \) is a connected graph.

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

Let \( x_i, x_{i+1}, ..., x_{\binom{n}{2}-1}, x_{\binom{n}{2}}, ..., x_{j-1}, x_j, x_{j+1} = x_i \) be a cycle in \( G \).

Suppose \( x_i \in R \). Then \( x_{i+1} \in W \), since the edge \( z_{i+1} = x_{i+1} - x_i \in B \), then \( x_{i+2} \in R \), since the edge \( z_{i+2} = x_{i+2} - x_{i+1} \in B \).

Continuing in this way, we see that if \( k \) is odd, then \( x_k \in W \), and if \( k \) is even then \( x_k \in R \). Since \( x_{j+1} = x_i \in R \), it implies that \( j + 1 \) is even, and thus the cycle is of even length. By theorem (2.8), then the graph \( G = (R, W, B) \) is a bipartite graph.

**Corollary(4.7):**

The set \( B \) of moves in theorem(4.4) is a Markov basis.
Proof:

Let $\mathcal{B} \subseteq \ker_\mathbb{C}(A)$ be a finite set of moves. From theorem (4.6) the graph $A^{-1}[t]_B$ is a connected graph. By definition (2.10) $\mathcal{B}$ is a Markov basis for $\mathcal{A}$. ■

Now, we will find the toric ideals that correspond to Markov basis for $2 \left(\frac{n}{2}\right) \times 2 \times \frac{n}{2}$ contingency tables.

**Corollary (4.8):**

Let $n$ be an even number, such that $n \geq 4$, and let $\mathcal{B}$ be a Markov basis for $\mathcal{A}$. Then the toric ideal for $2 \left(\frac{n}{2}\right) \times 2 \times \frac{n}{2}$ contingency tables in $A^{-1}[t]$, $I_\mathcal{A} = \langle p_i p_{n+j} - p_j p_{n+i} : i, j = 1, 2, ..., \frac{n}{2} \rangle$, such that $i \neq j \in \mathbb{C}[P_1, P_2, ..., P_n]$. ■

Proof:

From corollary (4.7), $\mathcal{B}$ is a Markov basis for $\mathcal{A}$, and by theorem (2.9) the set of binomials $\{p_{x^+} - p_{x^-} : x \in \mathcal{B}\}$ is a generating of toric ideal $I_\mathcal{A}$, and since $z_m \in \mathcal{B}$, $m = 1, 2, ..., k$, is a matrix of dimension $2 \times \frac{n}{2}$ and either it has two columns $(1, -1)'$, $(-1, 1)'$ and the other columns are zero, or it has two columns $(-1, 1)'$, $(1, -1)'$ and the other columns are zero. Then the toric ideal is the ideal $I_\mathcal{A} = \langle p_i p_{n+j} - p_j p_{n+i} : i, j = 1, 2, ..., \frac{n}{2} \rangle$, such that $i \neq j \in \mathbb{C}[P_1, P_2, ..., P_n]$. ■

**Example (4.9):**

For $n = 6$, according to theorem (4.2) there are 6 moves in a Markov basis $\mathcal{B}$ for $2 \times 3$ contingency table, and by remark (4.1)

$\mathcal{B} = \left\{\begin{bmatrix} 1 \\ -1 \\ 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 \\ 0 \\ -1 \\ 1 \end{bmatrix}, \begin{bmatrix} -1 \\ 1 \\ 0 \\ 1 \end{bmatrix}, \begin{bmatrix} 1 \\ 0 \\ 0 \\ -1 \end{bmatrix}, \begin{bmatrix} 0 \\ -1 \\ 1 \\ 1 \end{bmatrix} \right\}.$

By corollary (4.8) the toric ideal of $2 \times 3$ contingency table

$I_\mathcal{A} = \langle P_1 P_5 - P_2 P_4, P_4 P_6 - P_5 P_4, P_2 P_6, -P_3 P_9 P_2 P_4 - P_1 P_5 P_3 P_4, -P_1 P_6 P_3 P_5 - P_2 P_6 \rangle$. 


Remark (4.10): 

We can use corollary (4.8) to find the toric ideal \(2 \left(\frac{n}{2}\right) \times 2 \times \frac{n}{2}\) contingency tables without finding the Markov basis.

5. A New Module of Permutation The Pieces of Nucleotides in DNA Sequences

In this section, we construct a new model of genetic algorithm of permutation of the pieces of Nucleotides in aligned DNA sequences using an algorithm. Now we describe our model in the following steps.

Step (1):

Suppose we have \(m\) taxons of DNA sequences, each taxon of length \(L\) such as

| \(x_1\) | \(x_2\) | \(\ldots\) | \(x_n\) |
| \(\sum_{i=1}^{n} x_i\) |
| \(\sum_{i=2}^{n} x_i\) |
| \(\sum_{i=2}^{n} x_i\) |
| \(|\text{X}| = \sum_{i=1}^{n} x_i = L\) |

Taxon1: A G C T A A C GGTACGT···
Taxon2: CGATCTGAC CCGT T···
Taxonm: ACGTCA CGTA ACGC···

Then, we define a pattern \(i = i_1, i_2, \ldots, i_m\) 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 \(x_i\) is that \(i\) appears in our set of sequence data, and we refer to the number of frequencies by \(n\) where \(n\) is an even number greater than or equal 4.

Step (2):

We can input the patterns frequency \(x_i\) of above sequences in \(2 \left(\frac{n}{2}\right) \times 2 \times \frac{n}{2}\) contingency tables as follows:

Where \(|x| = \sum_{i=1}^{n} x_i = L\) is the length of sequences (the sample size), and \(x_1\) is the frequency of the first pattern.
\(x_2\) is the frequency of the second pattern.

\[x^n \frac{2}{2}\] is the frequency of the \(\frac{n}{2}\) pattern.

\(x^n \frac{2}{2+1}\) is the frequency of the \(\frac{n}{2}+1\) pattern.

\(x^n \frac{2}{2+2}\) is the frequency of the \(\frac{n}{2}+2\) pattern.

\(x_n\) is the frequency of the \(n\) pattern.

**Step (3):** Represent the contingency table \(x = \{x_i\}_{i \in I} \in \mathbb{N}^n\) as a \(n\)-dimensional column vector of non-negative integers \(x = (x_1, x_2, \ldots, x_n)'\), where ‘\(\)’ denotes the transpose of a vector or matrix, as in remark (2.2). Then \(x\) is a \(t\)-fiber (i.e \(x \in A^{-1}[t]\), where\(A^{-1}[t] = \{x \in \mathbb{N}^n: Ax = t\}\).)

**Step (4):**
\(A\) is \(\frac{n+4}{2} \times n\) matrix and From remark (4.3),

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

\(t\) is written as

\[
A = \begin{bmatrix}
x_1 \\
x_2 \\
x_3 \\
x_4 \\
\vdots \\
x_n \\
\end{bmatrix}
\]

\[
\begin{bmatrix}
t_1 \\
t_2 \\
t_3 \\
t_4 \\
\vdots \\
t_{\frac{n}{2}+2} \\
\end{bmatrix}
\]

Where \(Ax = t\) is written as

\[
\begin{bmatrix}
x_1 \\
x_2 \\
x_3 \\
x_4 \\
\vdots \\
x_n \\
\end{bmatrix}
= \begin{bmatrix}
t_1 \\
t_2 \\
t_3 \\
t_4 \\
\vdots \\
t_{\frac{n}{2}+2} \\
\end{bmatrix}
\]

where the columns of the matrix \(A\) index by the elements of the column vector \(x\).

**Step (5):**
Use remark (4.1), theorem (4.2), and corollary (4.7) to find the Markov basis \(B = \{z_1, z_2, \ldots, z_{\frac{n^2}{2}}\}\).

**Step (6):**
Use remark (4.5) and theorem (4.6) to find the connected graph \(G = (R, W, B) = A^{-1}[t]_B\), and \(2^{\frac{n}{2}}\) \((t - \text{fibers}) 2 \times \frac{n}{2}\) - contingency tables.
Step (7):

Use corollary (4.8) to find the toric ideals for \(2\left(\frac{n}{2}\right) \times 2 \times \frac{n}{2}\) - contingency tables.

Step (8):

Use the \(2\left(\frac{n}{2}\right)\) elements of the Markov basis \(B\) in step(6) to find the permutation of the pieces of nucleotides in aligned DNA sequences.

Example (5.1):

Suppose we have the following three aligned DNA sequences:

Taxon1: AGCTG A ATG GC
Taxon2: AGA TCTATC CA
Taxon3: A T TCA GACA AT

Step (1):

There are three taxons for above DNA sequences with \(|x| = \sum_{i \in I} x_i = L = 11\) and six patterns \(AAA, GGT, CAT, TTC, GCA, ATG, A\) \(AA, T TC, GCA, GCA,\) and \(CAT\) with frequencies \(2, 1, 2, 2, 3,\) and 1 respectively.

Step (2):

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

\[
\begin{array}{cccc}
2 & 1 & 2 & 5 \\
2 & 3 & 1 & 6 \\
4 & 4 & 3 & 11 \\
\end{array}
\]

Step (3):

Represent the contingency table \(x = \{x_i\}_{i \in I} \in \mathbb{N}^n\) as a \(n\) - dimensional column vector of non-negative integers \(x = (x_1, x_2, \ldots, x_n)\)' where' denotes the transpose of a vector or matrix, as in remark (2.2). Then \(x\) is a \(t\)-fiber (i.e. \(x \in A^{-1}[t]\), where \(A^{-1}[t] = \{x \in \mathbb{N}^n : Ax = t\}\)).
Step (4):

A is \( \frac{n+4}{2} \times n \) matrix and From remark (4.3),

Step (5):

Use remark (4.1), theorem (4.2), and corollary (4.7) to find the Markov basis \( B = \{ z_1, z_2, \ldots, z_{2(\frac{n}{2})} \} \).

\[
A = \begin{bmatrix}
1 & 1 & 1 & \cdots & 1 & 0 & 0 & \cdots & 0 & 0 \\
0 & 0 & 0 & \cdots & 0 & 1 & 1 & \cdots & 1 & 1 \\
1 & 0 & 0 & \cdots & 0 & 1 & 0 & \cdots & 0 & 0 \\
0 & 1 & 0 & 0 & \cdots & 0 & 0 & \cdots & 1 & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\
0 & 0 & 0 & \cdots & 1 & 0 & \cdots & 0 & 0 & 1^\frac{n+4}{2} \cdot n
\end{bmatrix}
\]

Where \( Ax = t \) is written as

\[
A = \begin{bmatrix}
1 & 1 & 1 & \cdots & 1 & 0 & 0 & \cdots & 0 & 0 \\
0 & 0 & 0 & \cdots & 0 & 1 & 1 & \cdots & 1 & 1 \\
1 & 0 & 0 & \cdots & 0 & 1 & 0 & \cdots & 0 & 0 \\
0 & 1 & 0 & 0 & \cdots & 0 & 0 & \cdots & 1 & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\
0 & 0 & 0 & \cdots & 1 & 0 & \cdots & 0 & 0 & 1^\frac{n+4}{2} \cdot n
\end{bmatrix}
\]

Step (6):

The connected graph \( A^{-1}[t]_B = G = (R, W, B) \) with \( 2(\frac{n}{2}) = 6 \)

(t- fibers) \( 2 \times 3 \)- contingency tables as vertices of it

Step (7):

By corollary (4.8), the toric ideal of \( 6 \times 2 \times 3 \)-contingency tables

\[
I_A = \langle P_2P_4 - P_1P_5, P_3P_4 - P_2P_6, P_1P_5 - P_2P_4, P_1P_6 - P_3P_4, P_2P_6 - P_3P_5, P_3P_4 - P_1P_6 \rangle.
\]

Step (8):

Then the permutation for the pieces of nucleotides in aligned DNA sequences under the set of Markov basis be \( B \) as follows:
Figure (3): The permutation of the pieces of nucleotides in aligned DNA sequences under the set of Markov basis.

**Remark (5.2):**

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

ii. We refer to \( ④ \) in the same figure to the hidden in the pattern frequency of DNA sequences.
Example (5.3):
Consider finding the connected graph for $6 \times 2 \times 3$--contingency table. Using the computer system Maple 13, the following commands will do the job:

```plaintext
> With( graph theory )
> G := (undirected, [(1, 2), (2, 3), (3, 4), (4, 5), (5, 6), (6, 1)])
> DrawGraph(G)
> IsBipartite(G)
> DrawGraph(G, style = circle)
> IsConnected(G)
```

True

> IsBipartite(G)
True

6. Conclusions
A Contingency table is a matrix of nonnegative integers with prescribed positive row and column sums. Contingency tables are used in statistics to store data from sample surveys. One of related problems for a survey of contingency tables is how to generate table from the set of all nonnegative $K_1 \times K_2$ integer tables with given row and column sums? (Fisher's Exact Test, Missing Data Problems).

A $K_1 \times K_2$--contingency tables $x$ can be written as a $n$--dimensional column vector of non-negative integers, where $n = K_1 K_2$, the contingency table $x$ can also be considered as an element in $t$-fibers $A^{-1}[t]$, where $A^{-1}[t] = \{ x \in \mathbb{N}^n : Ax = t \}$, $A : \mathbb{N}^n \rightarrow \mathbb{Z}^t$ is a linear transformation, and $t$ is $K_1 + K_2$--dimensional column
vector of row and column sums of the contingency table \( x \). Let \( B \) be a finite set, such that \( B \subseteq \ker A \) and let \( A^{-1}[t]_B \) be the graph with vertex set \( A^{-1}[t] \) and \( u - v \) an edge if and only if \( u - v \in \pm B \). If \( A^{-1}[t]_B \) is connected, then \( B \) is a Markov basis for \( A \). The question how to find a finite set \( B \subseteq \ker A \) such that \( A^{-1}[t]_B \) is connected ? is another problem, if \( B \) is found, then the set of binomials \( \{ p^{x^+} - p^{x^-} : z \in B \} \) generates the toric ideal \( I_A \) for \( A \) (i.e. \( B \) is used to find a toric ideal \( I_A \)).

In this Paper, we introduce an algorithm based on the actions of subgroups \( H_1 \) and \( H_2 \) of the dihedral group \( D_n \) on \( \ker A \) that gives solutions for all previous problems when \( n \) is an even number greater than or equal 4, \( K_1 = 2 \) and \( K_2 = \frac{n}{2} \). Remark(4.1), theorem(4.2), theorem(4.4), and remark(4.5) are used to generate \( 2 \left( \frac{n}{2} \right) \) contingency tables (\( 2 \times \frac{n}{2} \) matrices) in \( t \)-fibers \( A^{-1}[t] \), \( t = (t_1, t_2, t_3, ..., t_{\frac{n}{2}+2})' \), where \( t_1, t_2 \) are the row sums and \( t_3, ..., t_{\frac{n}{2}+2} \) are the column sums for each \( 2 \times \frac{n}{2} \) - contingency tables. Also, they used to find a finite set \( B \subseteq \ker A \). Remark(4.5) explains how to construct the graph \( A^{-1}[t]_B \) and theorem(4.6) states that \( A^{-1}[t]_B \) is a connected graph, this implies \( B \) is a Markov basis as mentioned in Corollary(4.7). Corollary(4.8) gives the toric ideal \( I_A \) for \( A \) using the Markov basis \( B \). All these results are used in section (5) to construct a new model of genetic algorithm of permutation the pieces of nucleotides in aligned DNA sequences.

**References**


[12] S. Aoki, and A. Takemura, "The list of indispensable moves of unique minimal Markov basis of $3 \times 4 \times K$ and $4 \times 4 \times K$ contingency tables with fixed two-dimensional marginals", Technical Report METR 03-38, Department of Mathematical Engineering and Information Physics, The University of Tokyo. Submitted for publication, (2003).