An algebra for biological sequences

Joyanta Pal, Dipankar Mazumdar, Swapan Raha
Department of Mathematics, Visva-Bharati
Santiniketan, West Bengal, India

ABSTRACT
In this paper, an attempt is made at an algebraic formulation of biological sequences. An algebraic structure is constructed for a given chromosomal string and segmentation. It is shown that this algebra represents, the most common chromosomal mutational mechanisms. Interpretation of the mathematical study is based on biological knowledge. Basic results are derived from the behavior of the chromosomal segments. This leads us to a new way of manipulating chromosomal mutation with mathematical and models.

Copyright © 2016 International Journal for Computational Biology. http://www.ijcb.in. All rights reserved.

How to Cite:

INTRODUCTION
A biological sequence is a sequence of characters from an alphabet. For DNA sequence, the alphabet is A, C, G, T, for RNA sequence, alphabet is A, C, G, U, and for protein sequence, the alphabet is A, R, N, D, C, Q, E, G, H, I, L, K, M, F, P, S, T, W, Y, V. The entire DNA [2] of a living organism is called its genome. Organisms are divided into eukaryotes and prokaryotes. Each cell contains the same genome. The DNA of eukaryotes is enclosed in a nucleus. Unlike the prokaryotes, eukaryotic genome is usually not a single string, but a set of strings called chromosomes. Chromosome is the unit of inheritance within the nuclei of all eukaryote cells. Eukaryotic chromosomes may differ in size, shape and composition of DNA, proteins and RNA, as well as in their number and redundancy. All these features are subject to evolutionary changes. Chromosomal mutations are variations in chromosome number or chromosome structure [1]. Chromosomal mutations affect the number and structure of chromosomes. Mutations, whether genetic or chromosomal are defined as any heritable change, and are the source of all genetic variations. Chromosomal abnormalities lead to diseases like cancer. Thus, an algebraic study on chromosomal mutation is of significance. In this paper, it is seen how biology motivates mathematics. However, a mathematical study of biology is also required.

Cells are fundamental working units of every living systems. All instructions needed to direct their activities are contained within the chemical DNA (Deoxyribo-Nucleic Acid). The letter from all organisms is made up of the same chemical and physical components. A DNA sequence is a particular arrangement of bases along the DNA string (e.g., ATCCGGACT). This order spells out the extra instruction required to create a particular organism with its own unique traits.

Biologists have shown keen interest in accepting tools and techniques from different disciplines of knowledge representation like mathematics, statistics and computer science for a better understanding of the biological
knowledge inherent in the data and realization of the secrets of life. This sequence analyses [3] should not be treated as an abstract task of string processing, because behind the string of symbols lies the whole complexity of molecular biology. This paper is the result of a multidisciplinary study, where mathematics combines with biology [4], to meet the demand for sophisticated analyses of biological knowledge.

The paper is organized into five sections. After a brief introduction in Section 1, an attempt is made to formally represent a biological sequence. Ultimately, in section 2, it has been shown that such formulation actually correspond to a non-commutative semi-group with identity. In section 3, different mutational activities have been represented as algebraic string manipulation. A brief report on further developments is presented in section 4. The paper is briefly concluded in section 5. A list of comprehensive references has been appended at the last.

2. STRUCTURE FORMULATION

A DNA sequence consists of four fundamentals of gene — adenine, thymine, guanine and cytocin symbolized as a, t, g, c. These constitute the alphabet [2] of DNA. Let us denote $S = \{a, t, g, c\}$ and $S^*$ as all possible sequences (that may be termed as string) formed by the members a, t, g, c, and $\Lambda \in S^*$, is the null element (i.e. null string). Clearly, $S^*$ is nonempty. Now, we define a binary operation [5] as the usual concatenation operation over the elements of $S^*$, i.e., if $x, y \in S^*$, then $x \ast y \in S^*$.

We define $x^0 = \Lambda$, $x^1 = x$, $x^2 = x \ast x$, and so on, for all $x \in S^*$. From the above formulation of operation and nature of $\ast$, we can say that $<S^*, \ast>$ forms a groupoid [6].

Biological Interpretation: The set $S^*$ is fundamental in a sense that most of its elements are either biologically predefined or have potential usage, any nucleotide sequence is a member of $S^*$ and $\forall x, y, z \in S^*$ we have the following;

1. $x \ast \Lambda = \Lambda \ast x = x$,
2. $y = z \rightarrow x \ast y = x \ast z$ and $y \ast x = z \ast x$,
3a. $x \ast y = x \ast z \rightarrow y = z$,
3b. $y \ast x = z \ast x \rightarrow y = z$,
4. $(x \ast y) \ast z = x \ast (y \ast z)$.

Clearly, the structure $<S^*, \ast>$ forms a non-commutative semi-group with identity.

Note 1: Imposition of different operations will enrich the structure. The set will be handled with rich mathematical structures. Here, we attempt to develop the mathematics relevant to biologists. The set $S^*$ has many elements that have no biological meanings. We would like to concentrate our study on parts that have some biological relevance. Accordingly, we need more specific restrictions on formation of the set $S^*$ so that it becomes more biologically relevant.

Chromosomes have different types of segments. They can be found to be useful in different dimensions. Some segments are genes, some are not. Genes have segments like introns and exons [7]. Tandem repeats are also chromosomal segments important for molecular or genomic study.

We know that a chromosomal string is a sequence of a, t, g, c i.e., a member of $S^*$. We consider a particular chromosomal string $C \in S^*$. Let $\Psi_C$ be the segmentation of $C$. Let $C = c_1 \ast c_2 \ast \ldots \ast c_n$ and consider $\Psi_C = \{c_1, c_2, \ldots, c_n\}$.

Example 1: Suppose,$\quad C = \text{atttgattacagataaagacagatagggagatagacagat}$.

Two possible segmentations are in figure

$$1\Psi_C = \begin{cases} 
\text{atttgattac} \\
\text{agataa} \\
\text{agacagatagggat} \\
\text{agacagat} \\
\text{atttgaa} \\
\text{ttgcatag} \\
\text{agataagac} \\
\text{a} \\
\text{gatagggagatagac} \\
\text{agat} 
\end{cases}$$

$$2\Psi_C = \begin{cases} 
\text{atttgattac} \\
\text{agataa} \\
\text{agacagatagggat} \\
\text{agacagat} \\
\text{atttgaa} \\
\text{ttgcatag} \\
\text{agataagac} \\
\text{a} \\
\text{gatagggagatagac} \\
\text{agat} 
\end{cases}$$
**Remark 1:** Other segmentations can also be made, depending upon specific requirements. If we are concerned with genes alone then the segments may be restricted to genes. On the other hand, if we are interested with parts of genes then the segments may be introns, exons etc. The length of the segments may vary from one nucleotide to thousands of nucleotides. A chromosome itself may be considered as a segment. The whole genome may be segmented. The segments then can be its parts such as chromosomes. For instance, let $C$ be the human genome. Then one $\mathcal{P}_C$ may be $\{c_1, c_2, \ldots, c_{46}\}$ where $\forall 1 \leq i \leq 46, c_i$ is a chromosome.

Human genes are broken into many exons, separated by seemingly meaningless pieces called *introns*. A gene broken into pieces may still produce the corresponding protein. To accomplish this, cells have to cut off the introns and all the exons be concatenated together. This happens in the *transcription* process.

Splicing (figure 1) is a modification of an RNA after transcription, in which introns are removed and exons joined. This is needed for typical eukaryotic messenger RNA before it can be used to produce a correct protein through translation.

**Example 2:** For a given chromosome $C$ of a specific species, the structure is represented in figure 2. A segmentation of this $C$ is,
Now, suppose for some specific case more filtration is needed. For example, on account of a particular reason we may need to study the exons only. In that case, only the exons will form a segmentation. Then,


There is no problem if this filtration is included in \( \Psi_C \) and without loss of generality, we can write \( \Psi_C^* \) instead of \( \text{exon}\Psi_C \), as,

\[ \Psi_C^* = \{ \text{spe5.chrom3.gen9.exo.27, spe5.chrom3.gen9.exo.28, spe5.chrom3.gen9.exo.29} \}

From the set \( \Psi_C \) we construct a set \( \Psi_C^* \) as in the following:

- \( \Psi_C^* \) is the space consisting of the non-empty set \( \Psi_C \) and the concatenation operator \(*\).

**Note 2:** Two elements \( x, y \in \Psi_C^* \) are adjacent if they occur adjacent to each other in \( C \).

This induces a structure on the elements of \( \Psi_C^* \) (as shown in figure 3).

![Figure 3: Structural representation of the relation over \( \Psi_C^* \).](image)

![Figure 4: Concatenation of the chromosomal segments \( x_1, x_2, \ldots, x_7 \) of (a) makes the segment (a). Deletion of segments \( x_2, x_3 \) and \( x_6 \) of (b) produce (c). The concatenation of the segments \( x_1, x_4, x_5, x_7 \) of (c) produce segment (d) \( \Psi_C^* \). We now define a relation adjacent to on \( \Psi_C^* \) as \( x \) is adjacent to \( y \) iff \( x \) and \( y \) are adjacent in \( \Psi_C^* \). Clearly, this relation is (a) reflexive (b) symmetric but not (c) transitive.\( http://www.ijcb.in \)
Example 3: In splice-sites (Figure 5) are the intron-exon junctions in the precursor mRNA of eukaryotes.

Most introns start from the sequence gt and end with the sequence ag (in the 5′ donor to 3′ acceptor direction), though the sequences at the two sites are not sufficient to signal the presence of an intron. Though the junctions are considered as sub-strings of length zero, they have special meaning. Like the exon-intron junctions that have been discussed, other types also carry biological importance. A biological interpretation is associated with a null segment.

**Figure 5: Splice sites**

**Definition 1:** The length of a string \( x \) is defined as the number of symbols present in \( x \in S^* \).

Clearly,

\[
\text{length}(\Lambda) = 0; \quad \text{length}(x * y) = \text{length}(x) + \text{length}(y); \quad \forall x, y \in S^*.
\]

**Deduction 1:** Thus, if \( x = x_1 * \Lambda * x_2 \), where \( x_1, x_2 \) are the segments then \( \text{length}(x) = \text{length}(x_1) + \text{length}(x_2) \).

**Theorem 1:** \(<S^*, *>\> is a semi-group.

**Proof:** To prove this, we have to show that * is closed and associative in \( S^* \). For this, let \( a, b \in \Psi^* \), \( a * b \in \Psi^* \) by the definition of \( \Psi^* \) and *.

Again, \( a * (b * c) = (a * b) * c \), \( \forall a, b, c \in \Psi^* \) as is clear from the operation of concatenation. Thus \(<\Psi^*, *>\> forms a semi-group.

**Note 3:** This semi-group is different from the semi-group \(<S^*, *>\> . The new one is more sophisticated as restricted by \( \Psi \), a biological factor. So its implication is more relevant to real life problem.

**Deduction 2:** If \( \Psi^*_c \) and \( \Psi^*_{c_0} \) are two semi-groups, then their Cartesian product is also a semi-group with respect to the same concatenation.

**Deduction 3:** The only idempotent of the semi-group \(<\Psi^*_c, *>\> is the null segment \( \Lambda \).

**Theorem 2:** \(<\Psi^*_c, *>\> is cancellative.
Proof: Since, * is cancellative and * is defined on $S^*$, which is cancellative, so $\langle \Psi^*_C, * \rangle$ is also cancellative.

Deduction 4: Let $\Psi^*_C = \{c_1, c_2, \ldots, c_{46}\}$. Then each $\langle c_i, \ast \rangle$, $1 \leq i \leq n$ is a cyclic semi-group. where $\langle c_i \rangle = \{c_i^0, c_i^1, \ldots, c_i^n, \ldots\}$

Now, if $c_i \in c_{\rho}$, then $c_i^j \in c_{\rho}$, for $j = 1, 2$...

Example 4: The repetition of string that occurs in DNA forms a cyclic semi-group. Tandem repeats occur in DNA when, a pattern of two or more nucleotides is repeated and the repetitions are directly adjacent to each other. As in the case of the string atcgattcgattcg, the sequence atcg is repeated three times.

Note 4: If 10 to 60 nucleotides are repeated, the string is called a mini-satellite, whereas those with fewer is known as micro-satellite or in short tandem repeat.

The order [6] of a cyclic semi-group is the total number of distinct strings present in the cyclic semi-group.

Regular Semigroup:

We know that a semi-group is regular [6] if for each $x \in S$, there exists $y \in S$ such that $x = xyz$

Now, by the construction of $\Psi^*_C$, and for each $x \in S$, there exits a $\Lambda \in \Psi^*_C$, such that $x = x \ast \Lambda \ast x$.

Note 5: So, we can take as $\Psi^*_C$ as a regular semi-group. In general, the identity element is not a necessary property of a semi-group, but we use it to develop $\Psi^*_C$ as a regular semi-group.

Now, we try to develop $\Psi^*_C$ in terms of classes.

Let us define a relation $\rho$ over $\Psi^*_C$. The relation $\rho$ holds between two elements of $\Psi^*_C$ if both of them have maximum same length of segments (in terms of supremum), that is,

$$a \rho b \text{ if and only if } \sup\{\text{length}(a_i); a_i \in a\} = \sup\{\text{length}(b_i); b_i \in b\} \text{ where } a, b \in \Psi^*_C.$$  

Then from the definition, $\rho$ is reflexive, symmetric and as well as transitive. So is an equivalence relation over $\Psi^*_C$.

Therefore, from the fundamental theorems on equivalence relation, $\rho$ can partition $\Psi^*_C$ in terms of disjoint classes and the set of all disjoint classes is denoted by $\Psi^*_C / \rho$.

3. MUTATION

There are different types of mutations on gene, namely,

(a) Inversion: An inversion (fig(10)) is a chromosomal mutation that results when a segment of a chromosome is excised and then reintegrated in an orientation of 180° (degrees) from its original position. When the inverted segment includes the centromere, the inversion [11] is called a pericentric inversion. Whereas, if the inverted segment occurs on one chromosome arm and does not include the centromere, the inversion is called a paracentric inversion. Genetic material is not lost when an inversion takes place, although there can be phenotypic consequences.
The above inversion mechanism induces an operation \( \sim \) — involution. The involution is a unary operator \( \sim : \Psi_c^* \rightarrow \Psi_c^* \) such that, \( \forall a, b \in \Psi_c^* \) 
\[ (a') = a \quad \text{and} \quad (a * b)' = b' * a' . \]

**U-Semi-group:**
In an semi-group, we define above unary operator such that \( \forall a, b \in \Psi_c^* \)
\[ (a') = a \], then this type of semi-group is called a U-Semi-group[6]. Clearly \( \Psi_c^* \) is a U-Semi-group.

**\( \ast \)-Semi-group:**
In a Semi-group, we define above unary operator such that \( \forall a, b \in \Psi_c^* \), \( (a * b)' = b' * a' \). Such a Semi-group is called a \( \ast \)-Semi-group[6]. Clearly, \( \Psi_c^* \) is a \( \ast \)-semi-group.
Note 6: $(\Psi^*_c, * , \cdot)$ is a monoid with involution.

Theorem 3: $\Lambda' = \Lambda$ for the identity $\Lambda$.

Proof: $\Lambda = \Lambda \ast \Lambda'$ (by the definition of the identity element, $x = e \ast x$).

$= (\Lambda' \ast \Lambda')'$ (by the first property of the definition of involution)

$= (\Lambda \ast \Lambda')'$ (by the definition of the identity element, $x = e \ast x$)

$= \Lambda$ (by the first property of the definition of involution)

(b) Deletion:

Definition 3: A deletion (fig(12)) is a chromosomal mutation involving the loss of a segment of a chromosome. The deleted segment may be located anywhere along the chromosome. A deletion [12] starts where breaks occur in chromosomes. These breaks may be induced, for example, by agents such as heat, radiation (especially ionizing radiation), viruses, chemicals, transposable elements or by errors in recombination. Because a segment of chromosome is missing, deletion mutations cannot revert to the wild-type state. A number of human disorders are caused by deletion of chromosome segments. One human disorder caused by deletion is cri-du-chat syndrome, which results from an observable deletion of part of the short arm of chromosome 5. Children with cri-du-chat syndrome are severely mentally retarded, have a number of physical abnormalities and cry with a sound like the mew of a cat (hence the name, which is French for ‘cry of the cat’).

Let us define a homomorphism $\Delta$ over $(\Psi^*_c, * , \cdot)$ such that

$\Delta (x) = \Lambda$.

for all $x \in (\Psi^*_c, * , \cdot)$ and $x$ is singleton that is one of $a, t, g, c$.

Figure 11: Deletion

Theorem 4: If, $a' = b'$ then $a = b$, $\forall a, b \in \Psi^*_c$.

Proof: $a' = b' \Rightarrow (a')' = (b')' \Rightarrow a = b$.

Theorem 5: If, $a' = b$ then $a = b'$, $\forall a, b \in \Psi^*_c$.

Proof: $a' = b \Rightarrow (a')' = b' \Rightarrow a = b'$.

Theorem 6: $(\Delta (a \ast b))' = \Delta (b' \ast a')$, $\forall a, b \in \Psi^*_c$. 

Figure 12: Forms of deletion
Proof: \((\Delta (a * b))' = (\Lambda)' = \Lambda\) and \(\Delta (b' * a') = \Lambda\).

**Corollary 1:** \((\Delta (a))' = \Delta (a' ), \forall a \in \Psi_c^*\).

**Theorem 7:** If \(a_1 * a_2 * \ldots * a_n = \Lambda\), \(\forall a_i \in \Psi_c^*, 1 \leq i \leq n\) then \(a_i = \Lambda, \forall 1 \leq i \leq n\).

**Proof:** The proof follows from the definition of the null string.

**Theorem 8:** If \(a' = \Lambda\), then \(a = \Lambda, \forall a \in \Psi_c^*\).

**Proof:** \(a' = \Lambda = \Lambda' \Rightarrow a = \Lambda\).

**Theorem 9:** \(a'i * a'j = a'i + j\)

**Proof:** \(a'i * a'j = a' * a'a'a'a'+ \cdot \cdot \cdot \cdot a' + j \) times = \(a'a'a'a'+ \cdot \cdot \cdot ' + j \) times = \(a'i + j\)

**Theorem 10:** \((a'i * a'j)' = b'i * a'j\)

**Proof:** \((a'i * a'j)' = (a' * a'a' \cdot \cdot \cdot a'a'+i \cdot \cdot \cdot a'+j \) times \(a'a'a'a'+ \cdot \cdot \cdot ' + j \) times = \(b'i * a'j\)

(c) **Insertion:**

**Definition 4:** Insertion is mutation in which extra base pair is inserted into a new place in a DNA.

---

(c) **Duplication:**

**Definition 5:** A duplication is a chromosomal mutation that results in the doubling of a segment of a chromosome. We consider a mapping \(D\) over the same semi-group by

\[ D(x) = x * x; \forall x \in (\Psi_c^*, \cdot, ') \]
D is called the Duplication operator (fig(16))[14].

Duplication has played an important role in the evolution of gene families. The size of the duplicated segment may vary considerably and duplicated segments may occur at different locations in the genome or in a tandem configuration (that is, adjacent to each other). When the order of genes in the duplicated segment is the opposite of the order of the original, it is reverse tandem duplication; when the duplicated segments are tandemly arranged at the end of a chromosome, it is a terminal tandem duplication.

**Theorem 11:** \((D(a \ast b))' = D(b' \ast a'), \forall a, b \in \Psi_c^\ast\).

**Proof:**
\[
\begin{align*}
(D(a \ast b))' &= ((a \ast b) \ast (a \ast b))' \\
&= (a \ast b)' \ast (a \ast b)' \\
&= (b' \ast a') \ast (b' \ast a') \\
&= D(b' \ast a')
\end{align*}
\]

**Theorem 12:** \(\Delta (D(a \ast b)) = D(\Delta(a \ast b)), \forall a, b \in \Psi_c^\ast\).

**Proof:**
\[
\begin{align*}
\Delta (D(a \ast b)) &= \Delta ((a \ast b) \ast (a \ast b)) \\
&= \Delta (a \ast b \ast a \ast b) \\
&= \Lambda \\
&= \Lambda \ast \Lambda \\
&= D(\Lambda) \\
&= D(\Delta(a \ast b))
\end{align*}
\]

**Theorem 13:** If \(D(a) = \Lambda\), then \(a = \Lambda\), \(\forall a \in \Psi_c^\ast\).

**Proof:**
\[
D(a) = \Lambda \Rightarrow a \ast a = \Lambda \Rightarrow a = \Lambda.
\]

**Theorem 14:** \(D(a^i) = a^{2i}\)

**Proof:**
\[
\begin{align*}
D(a^i) &= (a \ast a \ast \cdots a(i\text{ times}) \ast a \ast a \ast \cdots a(i\text{ times})) \\
&= D(2a^{2i}) \\
&= a^{2i}
\end{align*}
\]
Theorem 15: \((D(a^i))') = a^{2i}
\]
\[Proof: \ (D(a^i))') = (a^{2i})' = (a*a*...*a)' \ (2i \times) \ = a' * a' * ... * a' \ (2i \times) \ = a^{2i}
\]

Theorem 16: If \(D(a) = a\) then \(a = \Lambda, \forall a \in \Psi_c^*\).
\[Proof: D(a) = a \Rightarrow a*a = a \Rightarrow a*a = a*\Lambda \Rightarrow a = \Lambda \ (by \ left \ cancellation \ property) \]
Hence the proof.

Corollary 2: (a) If \(D(a_1 * a_2 * ... * a_n) = a_1 * a_2 * ... * a_n\) then \(a_i = \Lambda, \forall 1 \leq i \leq n, \forall a_i \in \Psi_c^*\).
(b) If \(D(a_1 * a_2 * ... * a_k * ... * a_n) = a_1 * a_2 * ... * a_k \rightarrow a_n\) then, \(a_i = \Lambda, \forall i \neq k \ and \ 1 \leq i \leq n\).
(c) If \(D(a) = a^i\) and \(i \neq 2\) then \(a = \Lambda\).
(e) Transposition: It is also a type of mutation.

Definition 6: The transposition [15] operator \(T\) is a composition of \(D\) and \(\Delta\) (fig(17)), and we write

\[T = \Delta o D,\]
where \(o\) is the ordinary composition of mapping (fig(18)).

4. FURTHER DEVELOPMENT—\(\sigma\)-ALGEBRA
For the set \(\Psi_c^*\), let us consider the collection of all subsets of \(\Psi_c^*\) and denote it by \(F\). Then \(F\) forms a sigma-algebra as it satisfies all the axioms of \(\sigma\)-algebra, that is,
(a) \(\Lambda \in F\),
(b) If \(A \in F\) then the complement \(A^c \in F\),
(c) If \(A_1, A_2, ..., A_n\) is a countable collection of sets in \(F\), then their union is also countable.

Note 7: (a) Clearly \(\emptyset, \Psi_c^*\) is a subset of \(F\) and \(F\) is a subset of \(P(\Psi_c^*)\), where \(P(\Psi_c^*)\) is the power set [6] of \(\Psi_c^*\).
(b) We have already constructed the partition of \( \Psi_c \), then the set of all partitions form a \( \sigma \)-algebra.

Now we know that all subsets of \( \Psi_c \) is generated by

\[ B = \{ a, t, g, c \} \]

so, \( B \) generates this \( \sigma \)-algebra and it is denoted by \( \sigma \) (B).

**\( \pi \)-system**: We consider the set \( \Psi_c \), then \( F \) is a \( \pi \)-system as \( F \) is closed under finite intersection. That is if \( A, B \in F \), then their intersection belongs to \( F \).

So, \( F \) can be considered as a \( \pi \)-system[6].

**\( \lambda \)-system**: We can also take \( F \) as \( \lambda \)-system, as

(a) \( F \) contains the null string \( \Lambda \).
(b) \( F \) is closed under complement, i.e., if \( A \in F \), then \( A^c \in F \).
(c) \( F \) is closed under countable disjoint union, if \( A_1, A_2, ..., A_n \in F \) and are pairwise mutually disjoint, then their union belongs to \( F \).

So, \( F \) can be considered as a \( \lambda \)-system[6].

**Deduction 6**: By Dynkin \( \pi - \lambda \) theorem, we can conclude that \( \sigma \) (F) is a subset of \( F \).

5. CONCLUSION

Mathematical modelling is a key issue central to all disciplines of sciences — physical science or life science. It is used to analyze systems and helps gain a better understanding of the functioning of the system before it is actually implemented. They allows analysis and/or synthesis with a substantial degree of accuracy yet with low cost. It is true that only a part of the organisms’ behaviour could be modelled mathematically as the underlying world has rarely been subject to formal treatment of mathematical modelling. Biological systems are inherently complex. As the complexity of the systems’s behaviour increases the ability to make precise yet significant conclusion/inference on its behaviour diminishes and ultimately becomes mutually exclusive characteristics.

This research has attempted to present a systematic development of some mathematical tools, that have well developed formal foundations, which could be used to understand formally part of the biological activities viz., mutations and gene expressions in living organisms. These biological activities are first abstracted as strings of meaningful symbols/sequences. Two groups of researchers will benefit from this work. In the first place, mathematicians may find that certain mathematical tools are indeed applicable to expression of behaviour of living organisms. They will then be able to address several interesting open problems. Secondly, researchers involved in application of computational tools to represent and analyze a substantial biological data, having inherent biological complexity will be able to acquire a better understanding of the mathematical basis of such computational tools[18].

The research reported in this paper has attempted to develop an algebra for the manipulation of strings which are found as mathematical models of several biological activities in living organisms. Extensive study of chromosomal aberrations reveal that chromosomal mutations are variations in chromosome number or chromosome structure. Chromosomes are represented mathematically as a biological sequence of finite length. Chromosomes are conceived as composed of a number of segments that are biologically significant. An algebraic structure is constructed for a given chromosomal string and a given segmentation. This algebra represents the most common chromosomal mutational mechanisms in a formal manner. Interpretation of the mathematical study is based on biological knowledge. Some basic results are derived from the behaviour of the chromosomal segments. This will open a new avenue to researchers and make possible study of diseases, their prevention and cure. Different behavioural aspects of living organisms caused by chromosomal aberrations can also be studied theoretically. Some chromosomal mutational mechanisms are not included in the present study. Inclusion of these mechanisms will surely enrich the algebra. Other mechanisms however uncommon may also be included.

ACKNOWLEDGEMENT

This research has been partially supported by the UGC SAP (DRS) Phase-III project under the Department of Mathematics, Visva-Bharati.
REFERENCES