# Rule Based Matrix Insertion Deletion Scheme for Improved Bio Molecular Computing

**Kotteeswaran C, Khanaa V, Rajesh A**

*Abstract: The problem of RNA editing has been well studied and there are number of approaches discussed earlier but suffer to achieve higher performance. To improve the performance an rule based approach is discussed in this paper. The RNA sequence generally consists of DNA sequences being represented as chain of string values. The malformed sequence has been used to identify the presence of any disease and such malformation encourages the occurrence of any disease to be happen. The rule based approach reads the RNA sequence and verifies the presence of tiny sequences specified in the rule set. The presence of malformed sequence has been deleted using matrix insertion deletion and has been added to produce new sequence of RNA. The classification is performed based on the RNA Sequence Support measure (RSSM) being estimated towards different class of sequences. The proposed method improves the performance of bio molecular computing to support the disease prediction.*
*Keywords: DNA, RNA, Bio Molecular Computing, RSSM*

## I. INTRODUCTION

The modern society suffers with lot of unknown diseases and the number of diseases identified is getting increased every year. For any disease to be occur in the human body is highly supported by the DNA of the person. When a particular part of DNA-RNA sequence gets mutated, then particular disease being occurred. DNA represent the protein molecules of the human cells and RNA represent the DNA in form of string. By monitoring the RNA sequence of any person, the method can generate a new sequence by applying the molecular computing.

Bio molecular computing is the process of identifying the class of RNA sequence and verifying the RNA sequence for the membership to the class. However the sequence has no restriction for the size, the particles of the sequence will always follow a rule. The molecules can be classified into three classes in particular and they are represented by the alphabets. For example, the class A covers the alphabets A,B,D and B covers C,G,H,I where the alphabet C covers the rest of the molecules. So that, all the molecules can be covered within these three class. Similarly, any class of sequence would contain number of types and they can be represented as rules. Consider, the class X, which can contain N number of rules where each rule Ri{X} would be a sequence framed by the three classes in certain combination.

In this case, Ri{X}={ACBA} represent the single rule for the class X where each gene class of alphabet would contain different member of the class. For example Ri{X}={ACBA="D,X,G,A"}, where D is from class A, X is

**Revised Manuscript Received on** .

**Kotteeswaran C**, Research Scholar, Department of Computer Science and Engineering, BHIER, Chennai, Tamilnadu, India. Corresponding Author E-mail : kottees@yahoo.co.in

**Khanaa V**, Professor & Dean, Department of IT, BHIER, Chennai, Tamilnadu, India. E-mail: drvkannan62@yahoo.com

**Rajesh A**, Principal, C.Abdul Hakeem College of Engineering &Technology, Vellore, Tamilnadu, India. E-mail: amrajesh73@gmail.com.

from class C,G is from Class B and A is from Class A. Similarly, N number of rules can be defined for any class of sequences. As of the RNA sequences can be classified into number of classes, each class can maintain N number of rules and they have no restriction for their size and numbers.

The rule based matrix insertion deletion system applies the rules of the class on the input RNA sequence. If any of rule gets matched with the RNA sequence, the it will be ignored and if it does not gets matched, then particular tiny sequence would be deleted. Similarly, when the most part of the sequence gets matched then it will be modified and inserted into the matrix system. By doing so, the number of sequences can be improved and would be helpful to perform molecular computing.

The RSSM (RNA sequence support Measure), represent the support contained by the sequence considered towards a class. When the value of RSSM is higher than a threshold, the RNA sequence has been labeled with the class selected. Once the label has been assigned, then matrix insertion and deletion can be performed. The brief details of the proposed Rule based approach is discussed below.

## II. RELATED WORKS

The problem of RNA sequence insertion deletion has been performed by different methods. Such methods are discussed in this section.

In [1], the author consider the problem of ambiguity in insertion and deletion. Based on the uncertainty of ambiguity issues, the author defines three different complexity measures like Tlength-ins for insertion, Tlength-del for deletion and Tlength-con for concatenation. Similarly in [2], an complexity measure for concatenation has been presented for the insertion and deletion system. The method consider the problem of length in concatenation of the strings.

In [3], the author presented an introduction to quantum theory towards computation of atoms and cells. The author motivated to get the information to the reality problem and make the search result to be more associative. In [4], the author performed a detailed review on RNA rewriting based on trypanosome kinetoplasts. The domain of expressive power has been used to model the RNA strings as hierarchical way. Using the hierarchy the RNA string has been rewritten

The grammatical methods have been considered towards the problem of computer vision and pattern recognition in [5]. The grammars would have used to represent the RNA sequences in such a way that others can understand. Further, by semantic representations of the patterns would have supported RNA rewriting efficiently. The author presented a model how the RNA sequences can be represented in form of semantic manner. The RNA rewriting can be approached with the finite state machines

and how they can be performed is discussed in [6]. The author discussed a state based gesture learning and recognition model which uses spatial clustering which consider each as spatial temporal sequences. The method uses the color features of the human and build the state machine based on the features.

Towards the learning of activity grammar, an model has been presented in [7]. The method identifies the symbols and activity grammars using description length with least length. The fitness of data has been evaluated using the MDL measure.

In [8], the author presented a grammer based approach for the activity recognition. The method uses context free grammer which enables the definition of human activity. The same has been used for the representation of the RNA sequences. Similarly in [9], an probabilistic model is presented towards the detection and recognition of different agents. The method uses the grammars for the detection and recognition. The RNA sequences has been detected for the class based on the grammars.

In [10], the author presented a recombination of DNA using the pathway of RNA sequences. The method consider the integrity of chromosomes to repair the lesions on DNA sequences. The breaks, gaps and interlocks are identified in DNA sequences and has been introduces by E-coli bacteria. The replication of DNA has been identified and rectified using the method.

In [11], an purification method for E-coli has been presented. The method identifies the three different proteins of E-Coli. The interaction between the proteins has been identified and deleted.

All the above discussed methods suffer to produce efficient results on classification and produces poor accuracy.

## III. RULE BASED MATRIX INSERTION DELETION SCHEME

The proposed rule based approach receives the input RNA sequence and splits the sequence into number of sub strings in different levels and lengths. The generated RNA sequence set has been used to estimate the RNA sequence support Measure (RSSM). Based on the RSSM value a single class has been identified. Using the rule presence in the rule set, the matrix insertion and deletion has been performed. The detailed approach is presented in this section.

The block diagram of the rule based matrix insertion and deletion system has been presented in Figure 1. The Figure 1, also displays the various functional stages involved.
RNA Sequence Generation:
The RNA sequence is the chain of alphabet which can be of any length. First, the input sequence has been split into number of substrings. Consider the RNA sequence Rs of length X, then the method generates the sequence from the index 1 to X in the varying length from 1 to X. if the substring length is n where n is less than X, then n number of characters from the index has been copied to produce the sequence. By varying the index and substring length, a number of combinatory sequence can be generated. The RNA sequence algorithm works based on this and produces lot of sequences to perform further operations
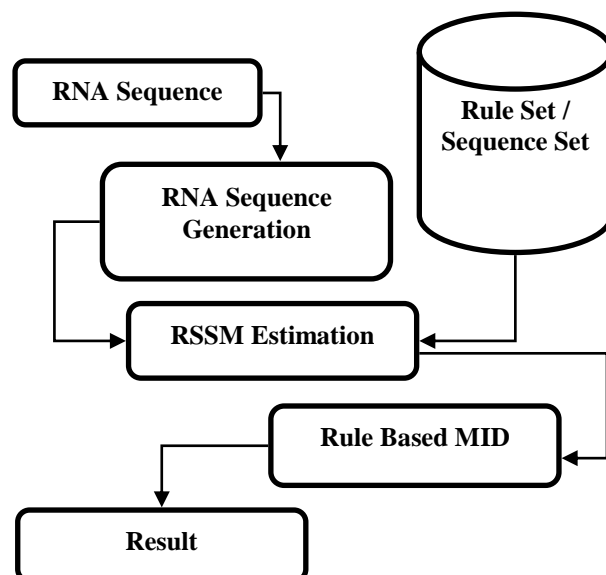


Figure 1: General Block Diagram of Rule Based MIDS

### 3.1 RNA Sequence Generation Algorithm
Intake: RNA-Sequence Rs.
Outcome: Sequences-Set Ss
Init
  Read RNA sequence Rs.
  Compute length of RS.
  Rlen = size(Rs)
  Initialize index to 1.
      Initialize sequence length to 3.
  for each index
Generate sequence s =
$$\int_{i=1}^{Rlen} \int_{j=3}^{Rlen-3} Substring(Rs, index, j)$$
Add to sequence set Ss = $\sum(Sequences \in SS) \cup S$
  End
Stop

The above discussed algorithm generates number of tiny sequences and has been used to estimate the RSSM value.
RSSM Estimation:
The RSSM measure represent the support for the membership of the sequence towards any class of sequences. It has been measured using the sequence set generated. For each sequence from the set, the method estimates the sequence frequency and inverse sequence frequency towards each RNA classes available. Based on the SF and ISF values, the method estimates the support for the sequence towards any class.

### 3.2 RSSM Estimation Algorithm
Intake: Sequence set Ss, Class Sequence Set Css.
Outcome: RSSM
Start
  Read SS and Css.
  For each sequence si from SS
Compute sequence frequency SF =
$$\sum_{i=1}^{size(Css)} \sum Css(i) \in si \Big/ size(Css)$$

Compute Inverse Sequence Frequency ISF =

$$\sum_{i=1}^{size(OCss)} \sum OCss(i) \in si \Big/ size(OCss)$$

OCSS – Other class sequence set.

End

$$\text{Compute RSSM} = \frac{\sum_{i=1}^{size(ss)} SF(si) \times ISF(si)}{size(Ss)}$$

Stop

The above discussed algorithm computes the RSSM measure based on the sequence frequency of tiny sequences. The estimated measure has been used for the classification finally.

### 3.3 Rule Based MID:

In this stage, the method reads the RNA sequence and generates the sequences. Using the sequence generated, the method estimates the RSSM measure for each class available. Based on the value of RSSM a single class has been selected. With the selected class, the method reads the rule set, and for each rule available, the method verifies the completeness with the sequence given. If any of the rule is identified as gets matched, the possible sequences with the molecules of the specific class has been added to the matrix system and insertion is performed. If the sequence contains set of matches from the other class rules, then matrix deletion is performed.

### 3.3.1 Rule Based MID Algorithm:

Intake: RNA sequence Rns, RNA Matrix Rm, Rule Set R
Outcome: RNA Matrix
Start
  Read RNA sequence Rns.
  Sequence set Ss = RNA Sequence-
          Generation(Rns)
  For each class c
 RSSM = Compute RSSM(Ss,Class Sequence Set).
 End
Class C = Choose the class with maximum RSSM.
Read Rule set R.
For each rule Ri
Compute Rule completeness measure $R_{cm}$ =
$$\frac{\sum_{i=1}^{size(R)} R(i) \in \rightarrow Rns}{size(R)}$$
Compute $R_{cm}$ with other class Rule $OR_{cm}$ =
$$\frac{\sum_{i=1}^{size(OCR)} OCR(i) \in \rightarrow Rns}{size(OCR)}$$
If $R_{cm}$>Th and $OR_{cm}$<Th then
  Sequence set ss =
  Generate Possible Sequences(Rs,Ri)
 RSSM =
  $\sum(\text{Sequences} \in Rm) \cup (\sum \text{sequences} \in ss)$
Elseif $R_{cm}$>Th and $OR_{cm}$>Th then
  Perform matrix deletion.
   End
End
Stop

The above discussed algorithm estimates the rule completeness measure and RSSM values to perform matrix insertion and deletion.

### IV. RESULTS AND DISCUSSION

The proposed rule based matrix insertion and deletion scheme has been validated for its performance in different measures

and compared with the performance of various other methods. The method has been simulation using Matlab by considering the RNA data set with number of classes. The proposed method has improved the performance of matrix insertion and deletion to support the bio molecular computing. The result produced by the proposed Rule based approach is presented below.

The performance on insertion and deletion has been measured and compared with the results of other methods.
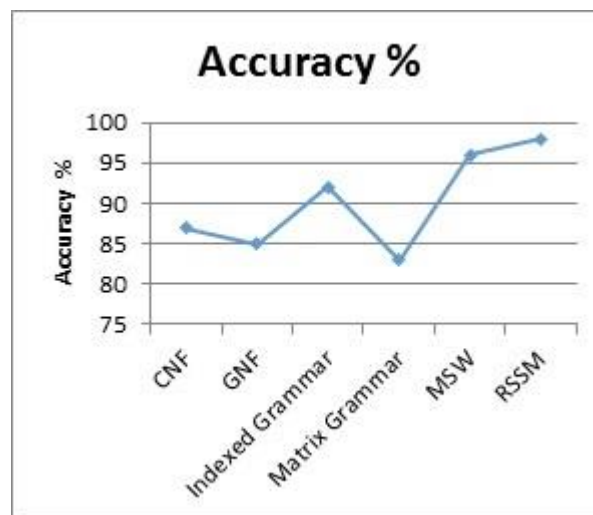


Figure 2: Comparison on Accuracy

The comparative performance has been presented in Figure 2. According to the values of Figure 2, the proposed RSSM algorithm has produced higher performance in insertion deletion than other methods
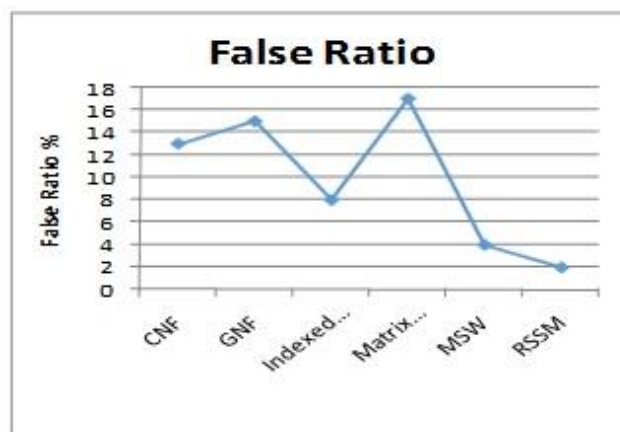


Figure 3: Comparison on False ratio

The performance on false ratio has been measured for the proposed RSSM algorithm and has been compared with the false ratio produced by other methods. The Comparative result has been presented in Figure 3, which shows the proposed RSSM algorithm has produced less false ratio than other methods.

The performance on matrix insertion deletion is depends on the performance of the classification. The classification performance has been measured for all the methods and compared with others. The result shows that the proposed RSSM algorithm has achieve higher classification performance than other methods.
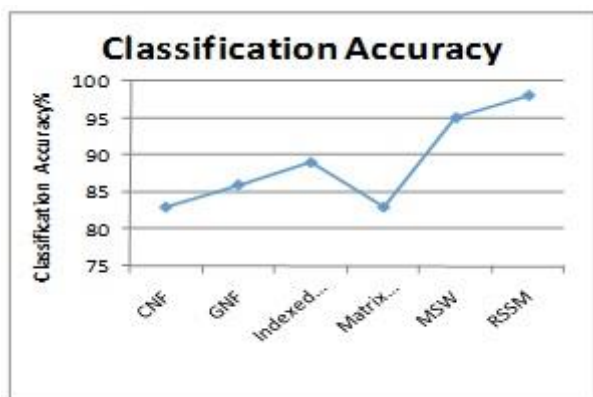
Figure 4: Comparison on classification accuracy

The risk detection accuracy has been evaluated and the proposed method has produced higher results on risk detection accuracy.
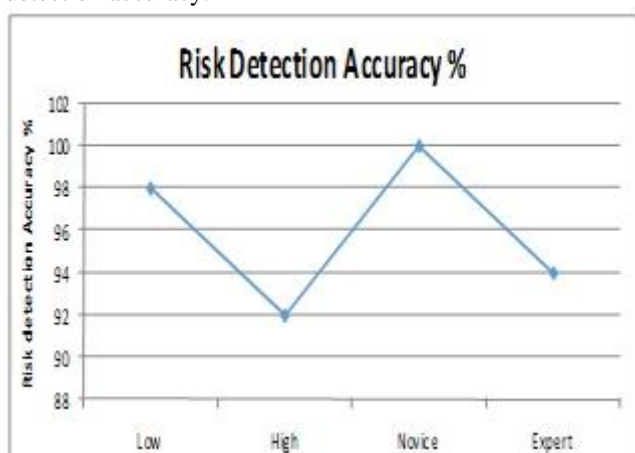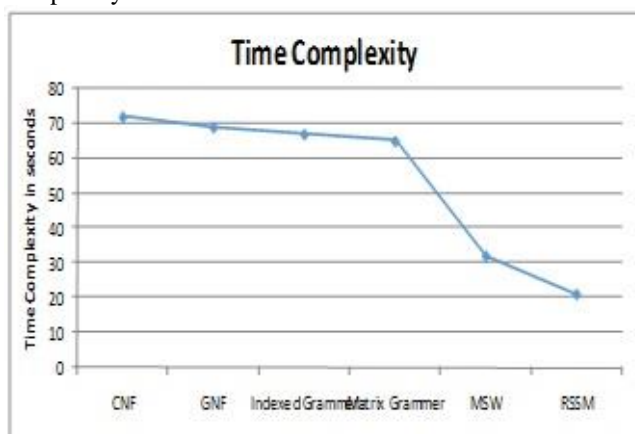


Figure 5: Comparison on risk detection accuracy

The performance on time complexity has been measured and compared with the result of other methods. The result shows that the proposed RSSM algorithm has produced less time complexity than other methods.



## V. CONCLUSION

A rule based matrix insertion and deletion system is presented. The method reads the input RNA sequence and generates the sequence set in varying index and varying length. For each sequence the method estimates the RSSM measure based on which the method identifies the class of RNA. Using the rule set of the selected class, the method estimates the RCM and ORCM measures to perform matrix insertion and deletion operations. The inclusion of RSSM has improved the performance of matrix insertion deletion system with higher accuracy.

## REFERENCES

1. Krithivasan. K "On the ambiguity and complexity measures in Insertion-Deletion Systems", in International Conference On Bio-Inspired Models of Network, Information, and Computing Systems (pp. 425-439) Springer, Berlin, Heidelberg. 2010.
2. Sergey Verlan. "On minimal context-free Insertion-Deletion Systems", Journal of Automata, Languages and Combinatorics, Vol 2 Pp. 317–328, 2007.
3. Cristian S. Calude "Computing with cells and atoms: An introduction to Quantum DNA and Membrane Computing, CRC Press (2001)
4. F. Biegler, "Regulated RNA rewriting: Modelling RNA editing with guided Insertion",Theoretical Computer Science, Vol.387, Issue 2: Pp 103-112, 2007
5. Chanda, G "Grammatical methods in computer vision: An overview", Technical Report, Georgia Institute of Technology, 2004.
6. Hong, P "Gesture modeling and recognition using finite state machines", In automatic face and gesture recognition, Proceedings. Fourth IEEE International Conference on, pp:410–415 , 2000
7. Kitani, K.M " An mdl approach to learning activity grammars" in proceedings of the korea-japan joint workshop on pattern recognition vol. 106, no. 376, pp. 19-24)
8. Ryoo, M. S "Recognition of composite human activities through context-free grammar based representation", In Computer Vision and Pattern Recognition, IEEE Computer Society Conference on, volume 2, pp:1709–1718,2006.
9. Ivanov, Y " Recognition of visual activities and interactions by stochastic parsing" IEEE Transactions on Pattern Analysis and Machine Intelligence, vol.22, issue8: pp:852–872,2000.
10. Spies, M. "Homologous recombination by RecBCD and RecF pathways", in the bacterial chromosome (pp. 389-403). American society of microbiology, 2005.
11. Shereda, R.D "A central role for SSB in escherichia coli RecQ DNA helicase function", Journal of Biological Chemistry, Vol.No 282, Page: 19247–19258, 2007.