Real Time RNA Sequence Edition with Matrix Insertion Deletion for Improved Bio Molecular Computing using Template Match Measure

Kotteeswaran C[1]
Research Scholar,
Department of Computer Science and Engineering,
Bharath University, Chennai, Tamilnadu, India

Khanaa V[2]
Professor and Dean,
Department of Information Technology
BharathUniversity, Chennai, Tamilnadu, India

Rajesh A[3]
Principal
C.Abdul Hakeem College of Engineering and Technology
Melvisharam, Tamilnadu, India

***Abstract*—The RNA sequence editing has become a challenging task in the molecular computing. There are number of approaches that have been discussed earlier for the problem RNA editing in bio molecular computing, but they suffer to achieve higher performance. To improve the performance, an real time approach has been presented which uses sequence depth measure (SDM). The method receives the RNA sequence and estimates the depth measure for different sub sequences generated. Based on the SDM value, a cumulative sequence match measure (SMM) has been measured to classify the sequence towards the different classes available. The matrix insertion and deletion is performed based on the template match measure (TMM) which has been computed based on the matches found in the templates available for different classes. The experimental results of our approach prove to outperform in terms of Accuracy, Risk Detection Accuracy, Time Complexity and False Classification Ratio which in turn increases the performance of bio molecular computing and matrix insertion deletion.**

***Keywords*—Bio Molecular Computing; RNA Sequence; SDM; SMM; Templates; TMM**

I. INTRODUCTION

The modern society has higher influence on new type of diseases which cannot be predicted. Every day, the researchers found new type of diseases appear in the human anatomy. There are number of researches on going to identify the solution and cause for the disease identified. However, for any disease to be occur on human body has higher support of their DNA system. The gene present in the human body only supports the arrival of the disease. For example, when you look at the DNA sequence of cancer affected peoples, you can identify the presence of certain sequence in all the patients DNA sequence. So the particular sequence encourages the disease and by identifying the sequence and modifying they would reduce the risk of getting into the disease.

The bio molecular computing is the presence of analyzing the RNA sequence which has been generated from the DNA pattern towards different disease classes. The bio molecular computing has been carried out towards finding the solution for different diseases. In general, the RNA sequence is the collection of chain of proteins and molecules. By analyzing the sequence, first we can identify whether the sequence is complete or not and also it has no restriction for their length. However, the sequence should follow a pattern or template which represents the membership of the set of sequences. In most cases, the RNA sequence has to be identified for their incorrect sequence and has to be removed and modified to produce a new sequence called RNA editing.

To perform RNA editing, there are number of approaches available and the matrix insertion deletion system is one which represent the RNA sequence in form of matrix. On the matrix available, according to certain forms and rules, the new sequence can be added and the existing sequences can be updated or deleted for specific sequence identified as malformed. By adding and removing the sequence from the entire RNA sequence, the sequences belongs to the class can be tuned. Because the sequence has been used to identify the possibility of disease and to identify the sequence which has been encourages the disease, they has to be well classified. To perform such classification there are number of methods and measures available. This paper present a real time approach which uses the sequence match measure which has been measured based on the appearance of the gene sequences in the RNA sequences set of any class available.

To classify the RNA sequence, the sequence depth measure is used which has been computed based on the depth of similarity available between the tiny sequence and the entire sequence set. By computing the SDM measure, the depth of closure can be identified for each class which has been used to perform classification. On the other side, the TDM measure has been used to perform matrix insertion deletion, which has been computed based on the templates of

sequences available for the class. The comprehensive explanation on our approach is presented in Section III.

II. Related Works

Variety of methods are available for the classification and editing of RNA sequences. This section discusses various techniques that are specifically related to the problem of RNA editing.

The vital measure of insertion deletion system that are widely used in the literature are ambiguity and complexity measures [1], Typically, we outline the many stages of ambiguity ($i = 0,1,2,3,4,5$) for insertion-deletion systems. Then, we attempt to exemplify that there are characteristically i-ambiguous insertion-deletion languages which are j-unambiguous for the various mishmashes $(i, j) \in \{(5,4), (4,2), (3,1), (3,2), (2,1),(0,1)\}$. Further, we prove an important result that the ambiguity problem of insertion-deletion system is undecidable. Finally, we define three new complexity measures $TLength - Con$, $TLength - Ins$, $TLength - Del$ for insertion-deletion systems and analyze the trade-off between the newly defined ambiguity levels and complexity measures.

On minimal context-free insertion-deletion systems [2], examine the type of context-free insertion-deletion systems. We know that such systems are common if the size of the inserted/deleted string is minimum three. We indicate that if this size is constrained to two, then the acquired systems are not common. We illustrate the acquired class and we bring together a new complexity measure for insertion-deletion systems which allows a well enlightenment of the acquired results.

Working with cells and atoms, an outline to Quantum [3], tends toward formal models of computability, we shall avoid over much “real” information to do with physics and biochemistry. The common approach we approve is to look to reality (whatever this means; passively, but safely, for us reality is what we can find in books and call such) in search of data supports (hence data structures) and operations on these data structures. With these constituents we can state a process (in the form of an arrangement of moves among configurations describing states of a system) which, provided that an input and an output can be linked with it, is considered a computation. The complete machinery is customized as a computing system.

Structured RNA rephrasing: Modeling RNA editing with guided insertion [4], study a model of string rephrasing based on the refined RNA editing mechanism found in trypanosome kinetoplasts. We establish simple properties of three main alternatives of this model which we indicate to form a strict order in terms of dramatic power. We also present a method and software for simulating real biological RNA editing via this model and apply the theoretical results to suggest real biological constraints on this process.

Grammatical methods in computer vision: An overview [5], examine several approaches and applications that have used grammars for solving inference problems in computer vision and pattern recognition. Grammars have been valuable since they are automatically easy to understand, and have precise sophisticated illustrations. Their ability to model semantic interpretations of patterns, both spatial and sequential, have ready them extremely popular in the research community. In this paper, we attempt to give an overview of what syntactic methods exist in the literature, and how they have been used as tools for pattern modeling and recognition. We also define numerous real-world applications, which have used them with great success.

Movement modeling and recognition using finite state machines [6], propose a state-based approach to movement learning and recognition. Using spatial grouping and temporal alignment, each movement is defined to be an ordered sequence of states in spatial-temporal space. The 2D image positions of the centers of the head and both hands of the user are used as features; these are located by a color-based tracking method. From training data of a given movement, we first learn the spatial information and then group the data into fragments that are automatically aligned temporally. The temporal information is further integrated to build a finite state machine (FSM) recognizer. Each movement has a FSM corresponding to it. The computational efficiency of the FSM recognizers allows us to achieve real-time on-line performance. We apply this method to construct an experimental system that plays a game of "Simon Says" with the user.

An Minimum Description Length (MDL) method to learning activity grammars [7], recommend a new technique for finding the best subset of non-noise terminal symbols and acquiring the best activity grammar. Our method uses the MDL principle, to evaluate the trade-offs between model complexity and data fit, to compute the difference between the results of each terminal subset. The assessment results are then used to find a class of candidate terminal subsets and grammars that remove the noise and allow the discovery of the basic structure of an activity. In this paper, we present the validity of our proposed method based on experimentations with artificial data.

Recognition of multiple human activities through context-free grammar based representation [8], describes a general approach for automated recognition of complex human activities. The approach uses a context-free grammar (CFG) based representation scheme to exemplify complex actions and interactions. The CFG-based representation enables us to formally define complex human activities based on simple actions or movements. Human actions are categorized into three types: atomic action, composite action, and interaction. Our system is not only able to represent complex human activities formally, but also able to identify represented actions and interactions with high accuracy. Image sequences are processed to extract poses and movements. Based on movements, the system detects actions and interactions occurring in a sequence of image frames. Our results show that the system is able to represent complex actions and interactions naturally. The system was tested to represent and identify eight types of interactions: approach, depart, point, shake-hands, hug, punch, kick, and push.

Identification of visual activities and interactions by stochastic parsing [9], describes a probabilistic syntactic approach to the detection and recognition of temporally

extended activities and interactions between multiple agents. The basic idea is to divide the identification problem into two levels. The lower level detections are performed using standard independent probabilistic event detectors to propose candidate detections of low-level features. The outputs of these detectors provide the input stream for a stochastic context-free grammar parsing mechanism. The grammar and parser provide longer range temporal constraints, disambiguate uncertain low-level detections, and allow the inclusion of a priori knowledge about the structure of sequential events in a given domain.

Homologous recombination by RecBCD and RecF pathways [10], a more immediate function of homologous recombination has been accepted: namely, it is a tool for the maintenance of chromosomal integrity that acts to repair DNA lesions, both doublestrand DNA breaks and single-strand DNA gaps, produced during the sequence of DNA replication. The familiar connection between the processes of replication and recombination was initially appreciated in the life cycle of bacteriophage T4(56) and then later accepted as an important determinant of viability in bacteria (39,45). In T4 phage, recombination is linked to replication to generate a high yield of phage DNA; in Escherichia coli, recombination is linked to replication to allow its completion when interrupted by DNA damage, and also to initiate DNA reproduction in the nonexistence of origin function.

A fundamental part for SSB in Escherichia coli RecQ DNA helicase function [11], use an similarity refinement scheme to recognize three heterologous proteins that associate with Escherichia coli RecQ: SSB (single-stranded DNA-binding protein), exonuclease I, and RecJ exonuclease. The RecQ-SSB interaction is direct and is intervened by the RecQ winged helix subdomain and the C terminus of SSB. Interaction with SSB has significant well-designed consequences for RecQ. SSB stimulates RecQ-mediated DNA unwinding, whereas deletion of the C-terminal RecQ-binding site from SSB produces a variant that blocks RecQ DNA binding and unwinding activities, suggesting that RecQ identifies both the SSB C terminus and DNA in SSB.DNA nucleoprotein complexes.

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

III. TMM BASED REAL TIME RNA EDITING SCHEME

The proposed TMM based RNA editing scheme receives the input RNA sequence and generates number of gene sequence tuples from the RNA sequence given. For each sequence tuple generated, the method estimates the sequence depth measure within a class of RNA sequence. Finally a cumulative SMM measure has been assessed to classify the class of RNA sequence given. Second, the method estimates the TMM measure towards each gene tuple generated within the class identified and based on that the method performs the matrix addition and deletion operations. The detailed approach is discussed below:

Fig. 1 shows the architecture of the proposed SDM based matrix insertion deletion system and shows various components in detail.

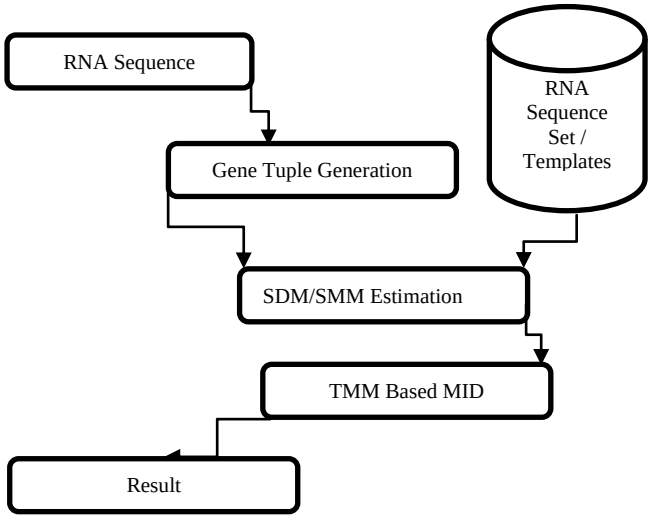

Fig. 1. Architecture of Proposed SDM based Matrix Insertion Deletion System.

A. Gene Tuple Generation:

The RNA sequences are varying in their length and they have no restriction for their size. The input RNA sequence has been read and the entire sequence has been split into number of tiny sequences. The sequences are generated in different length and from different position of the RNA string. The number of position considered is depending on the size of the string available. For each length available, the method generates all possible strings from the RNA sequence. The generated strings are added to the tuple set which has been used to estimate various measures.

Algorithm 1:

Input: RNA sequence S
Output: Tuple set Ts.
Start
  Read Rna sequence S.
  Compute overall length l = size(S)
  Initialize minimum length ml = 2.
  Initialize starting index si=1
  For each ml
    Generate tuple $Ts = \int_{i=si}^{l} \sum \text{sub string}(S, i, ml)$
  End
Stop

The above algorithm produces the tuple set from the input RNA sequence and has been used estimate different measures.

B. SMM Estimation

In this stage, the tuple set generated in the previous stage has been read. Using the tuple set and the RNA sequence set of different classes, the method estimates the SMM value. First, the method estimates the Sequence Depth measure (SDM) for each tuple available in the tuple set towards the sequence set of RNA. It has been measured based on the number of matches and appearances found in each sequence of the class. Finally, a sequence match measure (SMM) has been computed for the class based on the values of SDM of all tuples in the tuple set. Estimated SMM value has been used to perform classification.

Algorithm 2:
Input: Tuple set Ts, RNA set Rs
Output: SMM
Start
Read Ts, Rs.
For each Ti from Ts
Compute number of entries in each sequence $Ne = \sum_{i=1}^{size(Rs)} \sum Occurrence(Ti \in Rs(i))$
Compute number of matches in overall set $Nm = \sum_{i=1}^{size(Rs)} \sum Rs(i) \in Ti)$
Compute SDM = $\frac{Ne}{maxlength(\forall Rs(i))} \times \frac{Nm}{size(Rs)}$
End
Compute SMM = $\frac{\sum SDM}{Number\ of\ sequences\ of\ Rs}$
Stop

The above algorithm calculates the sequence match ratio for the period being considered and based on the sequence set of the period it has been estimated.

C. TMM Based MID

In this period, first the technique receives the RNA arrangement given. Using the RNA arrangement, the method generates the gene sequences as tuple set. Generated tuple set has been used to estimate the Sequence support ratio for different classes. Based on the arrangement support measure, the method selects a single class. For the selected class, the method estimates the TDM measure with the templates of strings available. Based on the TDM value, the method performs matrix insertion deletion.

Algorithm 3:
Input : RNA sequence R, Template set T, Matrix M, Tuple set Ts.
Output: Matrix M
Start
Read R, T, M.
Tuple set Ts =Generate gene sequence (R)
For each class c
SSM = Compute SSM(Ts, C)
End
Class c = Choose class with higher SSM.
For each tuple Tk from Ts
For each template Ti from T
Compute Number of Total match $Ntm = \sum_{i=1}^{size(T)} \sum T(i) \equiv Tk$
Compute number of partial match $Npm = \sum_{i=1}^{size(T)} \sum T(i) \approx Tk$
Compute TMM = Ntm×Npm
End
If TMM>Th then
Sequence set ss = Generate possible sequences with Tuple Tk.
Matrix M = $\sum(sequences \in M) \cup ss$
Else
Matrix M = $\sum(sequences \in M) \cap ss$
End
End
Stop

The above discussed algorithm estimates TMM measure for different template and based on the value the matrix insertion and deletion is performed.

IV. Results and Discussion

The proposed TDM based real time matrix insertion and deletion scheme has been implemented and evaluated for its performance. 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 method has produced the following results. The evaluation has been performed using various RNA data set and gene sequence set.

The accuracy on insertion deletion has been evaluated with different methods and presented in Fig. 2. The results show that the suggested method has generated greater precision than other approaches.

The risk detection accuracy has been evaluated and presented in Fig. 3. The comparative result illustrates that the suggested method has generated greater outcomes on risk detection accuracy.

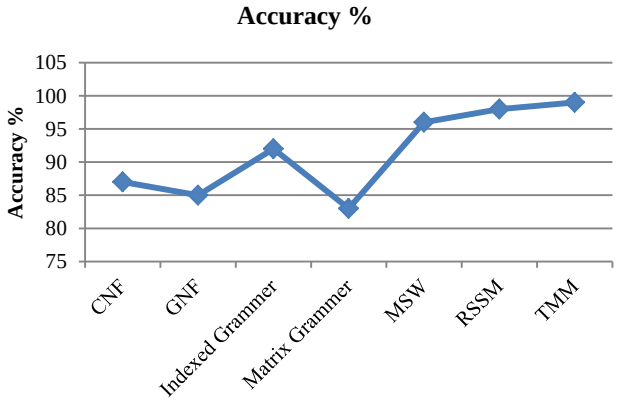

Fig. 2. Assessment on Accuracy.

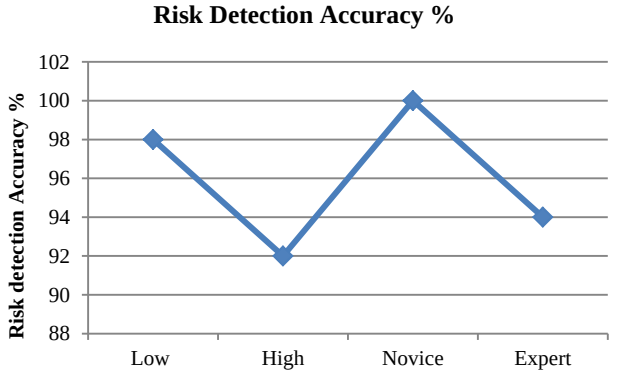

Fig. 3. Assessment on Risk Detection Accuracy.

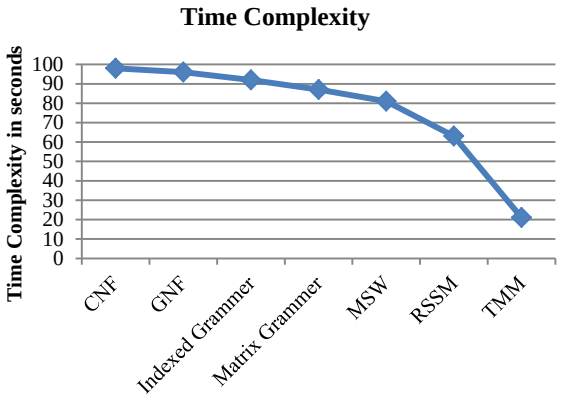

Fig. 4. Assessment on Time Complexity.

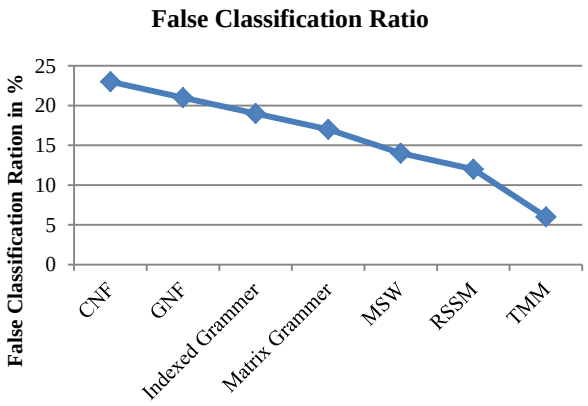

Fig. 5. Assessment on False Classification Ratio.

The time complexity being introduced by different method of insertion deletion systems has been measured and compared with the result of proposed TMM approach. The comparative result has been presented in Fig. 4. The result illustrates that the suggested TMM approach has lower time complexity than other approaches.

The false classification produced by different methods has been measured and compared with the outcome of suggested TMM procedure. The comparative result has been presented in Fig. 5 and shows that the proposed TMM algorithm has reduced the false ratio compared to other methods.

V. Conclusion

A real time template match measure based matrix insertion and deletion system is presented. The method reads the input RNA sequence and generates the tuple set based on the size of the RNA arrangement and the amount of indexes considered. Varying length of RNA sequence has been generated. With the generated tuple set, the method computes the sequence depth measure (SDM) and sequence match measure (SMM). The sequence depth measure represents the similarity of depth the sequence has with other one. Based on the SMM value the method performs classification, the class with higher sequence match measure has been selected and assigned with label. Based on identified class, the method estimates the template match measure (TMM) for different tuples available in the tuple set. Finally, according to the TMM value, the method performs matrix insertion and deletion operations. The matrix insertion is performed according to the TMM value and generates number of RNA sequences. The proposed method improves the performance of the system by improving the accuracy. Also, the method reduces the time complexity and false classification ratio.

References

[1] Kamala Krithivasan., Lakshmanan Kuppusamy., Anand Mahendran., Khalid M. “On the ambiguity and complexity measures in insertion-deletion systems”, LNCS, Proceedings of Bionetics, C 1–3, 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., Gheorghe Paun.: “Computing with cells and atoms, An introduction to Quantum”, DNA and Membrane Computing. Taylor and Francis, London (2001).

[4] F. Biegler, M.J. Burrell, and M. Daley “Regulated RNA rewriting: Modelling RNA editing with guided insertion”,Theoretical Computer Science, vol.387, issue 2: pp 103-112, 2007.

[5] Chanda, G., and Dellaert, F. “Grammatical methods in computer vision: An overview”, Technical report, Georgia Institute of Technology, 2004.

[6] Hong, P.; Turk, M.; and Huang, T. S. 2000, “Gesture modeling and recognition using finite state machines”, In Automatic Face and Gesture Recognition, 2000. Proceedings. Fourth IEEE International Conference on, pp:410–415. IEEE.

[7] Kitani, K. M.; Sato, Y.; and Sugimoto, A. “An mdl approach to learning activity grammars”, Technical Report 376, IEICE - The Institute of Electronics, Information and Communication Engineers, Tokyo, Japan, 2006.

[8] Ryoo, M. S., and Aggarwal, J., “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., and Bobick, A. Recognition of visual activities and interactions by stochastic parsing. Pattern Analysis and Machine Intelligence, IEEE Transactions on vol.22, issue8: pp:852–872,2000.

[10] Spies, M. and Kowalczykowski, S.C. “Homologous recombination by RecBCD and RecF pathways”, In The bacterial chromosome , pp. 389–403. ASM Press, Washington, DC, 2005.

[11] Shereda, R.D., Bernstein, D.A., and Keck, J.L., “A central role for SSB in Escherichia coli RecQ DNA helicase function”, J. Biological Chemistry, vol. 282: page: 19247–19258,2007.