Proposed Approach for RNA special regions prediction

Alaa Eldin Abdallah Yassin[1], Ahmed Farouk Al-Sadek[1],Amr Badr[2]

[1] Central Lab for Agricultural Experts Systems, Ministry of Agriculture and Land Reclamation
Giza, Egypt

[2] Department of Computer Science, Faculty of Computers and Information, Cairo University
Cairo, Egypt

Abstract

Untranslated region in RNA; Internal ribosome entry site (IRES) is crucial to exploring and deciphering the initiation mechanism of translation and replication of the virus, where IRES contains some conserved motifs existing in the function domain of IRES which are essential for IRES activity to start its. The existing RNA structure prediction programs are not sufficient in prediction IRES structure, because IRES have special nature differ from other RNA sequences. The objective of this paper is : Provide more reliability in the secondary structure level for RNA which have a special nature like function domain of IRES called; Apical. The proposed approach uses rule-based criteria, which use some predetermined parameters as energy, conserved motifs and loops of the predicted structure. The proposed approach applied on 50 sequences and have low error rate and up to 68 % accuracy at the first iteration and 100 % accuracy after the second iteration.

Keywords: IRES, Apical region, viral replication, FMDV, RNA prediction.

1. Introduction

Internal ribosome entry site (IRES) were first recognized about 30 years ago within the 5′ untranslated region of picornavirus [1]. In spite of the existence of IRES before the coded or translated region; in particular located at untranslated region of the virus; but it have an important role in virus replication process, because IRES contains some crucial domains which necessary to complete replication process [2,3,4]; where after the organism infected by the picornavirus, the viral genome is translated from IRES without competition with mRNA [11]. Biochemical studies demonstrate that IRES is a relatively long RNA region, exceed ~450 nucleotides that are predicted to fold in five or four domains depending on the virus family; and those domains not in the same level of importance, for example the central and functional domain in picornavirus domain 3 is, and domain 2 in Flavivirus [5]. Domain 3 in IRES consists of 2 regions, apical region and basal region, Apical region is the functional region which includes the conserved motifs and junctions which affect on the IRES in tertiary level [2, 4]. The conducted researches on IRES region whether at the first level (1D) [4,6,7] or in the second level (2D) [3,5] or in third level(3D) [2]; are not enough and need more efforts to well understanding the key of virus. Some tools and programs are developed to help in predicting RNA structure like MFLOD, where dynamic programming algorithm is reported for RNA secondary structure prediction by free energy minimization (MFE) [8,9], and this algorithm can be considered the most popular RNA prediction program specially after enhancement and updated versions, MFOLD web server can be used free through internet [10], RNA predictors also used widely Vienna package which are a collection of tools for folding, design and analysis of RNA sequences, one of those tools is used to predict RNA sequence named "RNAFold server", using MFE with partition function methodology [12], researchers also can access it freely via internet [13]. But the developed prediction tools and programs are not sufficient in predicting the special regions in viruses like IRES region in foot and mouse disease virus(FMDV) [11,12].So the proposed paper produced a new approach to set the optimal structure from all predicted structures which obtained from MFold tool, where as shown in figure 1 , MFold produce more than one structure to one RNA sequence. The proposed approach used rule-based criteria to score and select the most suitable structure of Apical sequence.

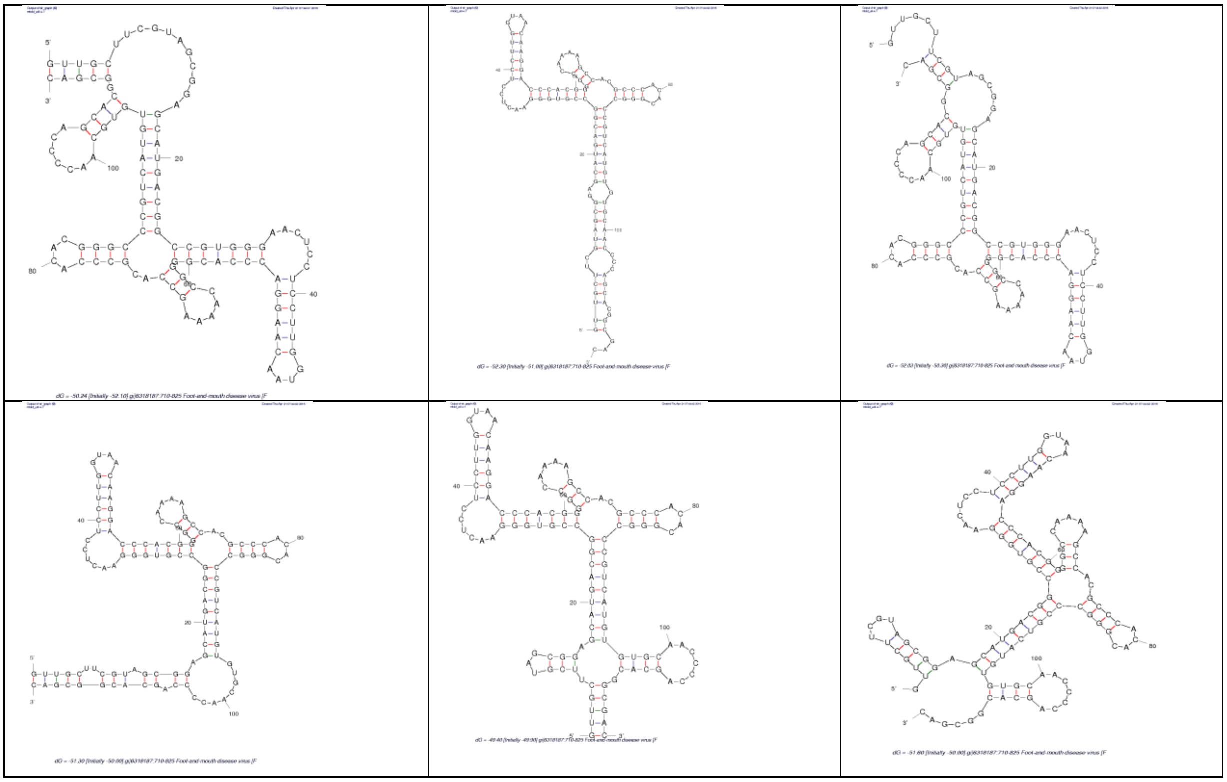
dG = -50.24 [Initially -52.10] gi|6318187:710-825 Foot-and-mouth disease virus [F
dG = -52.30 [Initially -51.00] gi|6318187:710-825 Foot-and-mouth disease virus [F
dG = -52.60 [Initially -50.30] gi|6318187:710-825 Foot-and-mouth disease virus [F
dG = -51.30 [Initially -50.00] gi|6318187:710-825 Foot-and-mouth disease virus [F
dG = -49.40 [Initially -49.90] gi|6318187:710-825 Foot-and-mouth disease virus [F
dG = -51.60 [Initially -50.00] gi|6318187:710-825 Foot-and-mouth disease virus [F

Figure 1:secondary structures of FMDV IRES: Apical, predicted using MFold web server . the predicted structures from A to F are predicted to only one sequence.

2. Methodology

As shown in figure 2, the proposed approach pass over several steps or phases to set the optimal structure of the target sequence . **First** we entered the sequence on MFold web server to obtain all possible structures to this sequence. **Then** enter each structure to the developed inference to apply rules to label this structure with one of the three cases (Accepted, May be or Not). **Then** pass to validation rules which select the near optimal structure of the target sequence. All previous steps will explained in the following sections.

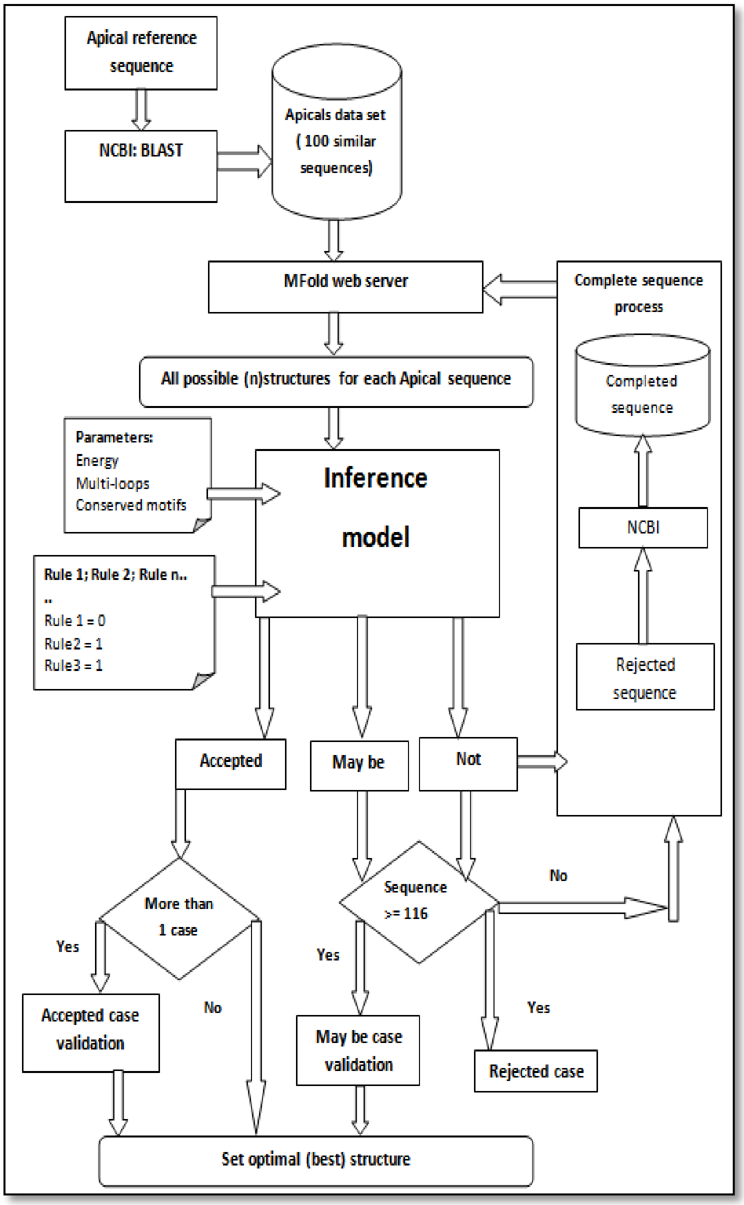

Figure 2: Flowchart for the proposed approach. RNA Apical sequence which collected previously, entered to MFold web server. RNA different structures are predicted. Enter each candidate structure to SASM model and apply rules to assign one from the three cases to the structure (Accepted, May be or Not). Then apply validation phase according to each case to set the optimal structure.

3. Predict possible structures for each sequence

In this phase we enter target sequence to MFLOD web server [10] , to predict all possible structures for this sequence, as presented in figure 1, and then we pass each candidate structure to the developed inference to apply its rules and produce the result of the fired rules. This process is shown if figure 2.

4. Inference

The developed inference contains a set of rules to label each candidate structure obtained from previous step. The suggested rules created according to some parameters or features determined with the help of the conducted biochemical studies on this point of research. The proposed approach use three different parameters; conserved motifs, multiple four-way junctions and energy. The three parameters will explained later and explain our hypothesis that the selected parameters or features of the predicted structure not at the same preference and we don't follow the general RNA prediction algorithms which give the energy of the predicted structure the most and may be the only way to prefer structure rather than structure, this hypothesis we mention in introduction section, and according to this hypothesis we give "conserved motifs" parameter the highest priority and "multi-branch loop" follow it , then we consider energy parameter. The suggested rules in the proposed prediction model are presented in figure 4.

4.1 Conserved motifs parameter

The conducted research on Apical region proved that this region contains two conserved motifs, GNRA and RAAA motif, where N is any nucleotides and R is A or G [2], so RAAA motif can be AAAA or GAAA ; and GNRA have eight probabilities (GAAA,GAGA,GGAA,GGGA,GCAA,GCGA,GUAA,GUGA), but some other research focus on these conserved motif and explained that the GNRA motif in FMDV virus almost is GUAA and located at the apex of the stem-loop motif [5, 19], and by observation RAAA motif in FMDV is AAAA. The proposed approach chose these conserved motifs to be the most important parameter in our proposed model because the biochemical studies proved that those motifs are critical for IRES function [20, 21] and affected in its stability and folded shape in the tertiary structure[16]. Also Apical region contains another motif name C-rich motif or loop, this motif is important in binding site region in tertiary structure phase [5], and by observation we found that C-rich motif is RACCCCR, where R is A or G. from previous words we conclude that the first parameter in the proposed model is Conserved motifs in the predicted structure, and these motifs are GUAA, RAAA and C-rich motif.

As shown in table 1 ; "conserved motifs" parameter has three cases, the best one is the high case which have value 3, which mean that the predicted structure contains the thee conserved motifs(GNRA,RAAA and C-rich), medium case has value 2 mean that the predicted motif contains two motifs from the three, and the last case is the low case which mean that the predicted motif contains only one conserved motif .

4.2 Four -way junctions parameter

As we mentioned before that Apical region contains multiple four-way junctions that formulate the shape of apical and IRES at all [2]. As shown in figure 3, apical region in FMDV contains two multi-branch loops or two four-way junctions, so we chose the existing of four-way junctions to be the second parameter in the proposed model. As shown in table 1 ; The second parameter "multi-branch loop" also has three different cases, the best one is high case with value equal to 2, which mean that this predicted structure contains two four-way junctions or two multi-branch loop in its folded shape, second case for this parameter is medium case with value one, which mean that this predicted structure contains only one 4-way junction, and the last and worst one is low case with zero value which tell us that the predicted structure don’t include any four-way junctions.

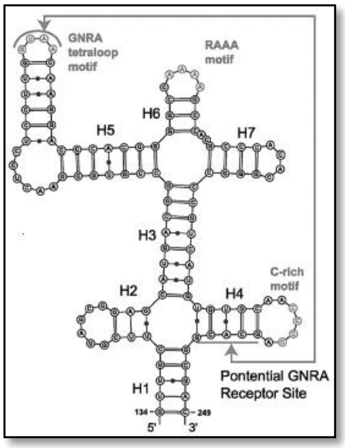

Figure 3:Apical region of FMDV in 2D, including the three conserved motifs and two multi-branch loops [2].

4.3 Energy parameter

The last parameter in the proposed model is Energy, because the energy is the main key features in RNA prediction programs, and the most famous RNA prediction program [10,13] use minimum free energy (MFE) as a methodology to its prediction, so we include energy in our work but give it low priority than conserved motifs and junction because the special case of RNA region that we want to predict. By observation of the energy value for the candidate structures obtained from MFold web server, we found that the values can’t be low than minus 60 and can’t be more than minus 40 for the different Apicals in the data set, and the best case here is the low case which mean that this structure folded with low energy and this hypotheses is reverse to MFold with order candidate structure according to the low energy value, see table 1.

Rule 1: if Conserved motifs is high and Multi-branch loop is high then case is (Accepted).
Rule 2: if Conserved motifs is high and Multi-branch loop is medium and Energy is low or medium then case is (May be).
Rule 3: if Conserved motifs is medium and Multi-branch loop is high then case is (May be).
Rule 4: if Conserved motifs is medium and Multi-branch loop is medium and Energy is low then case is (May be).
Rule 5: if Conserved motifs is medium and Multi-branch loop is medium and Energy is high or medium then case is (Not).
Rule 6: if Conserved motifs is low then case is (Not).
Rule 7: if Multi-branch loop is low then case is (Not).

Figure 4: The suggested rules of the proposed approach .

Table 1: Model parameters and its values

Parameter	High	Medium	Low
Conserved motifs	3	2	1
Multi-branch loop	2	1	0
Energy	<= 45	50<>45	60<>51

5. Applying rules

As shown in figure 2, all candidate structures for each Apical sequence that produced from prediction process through MFold web server, are entered to inference to apply suggested rules on each candidate structure and assessment this structure if it is “Accepted” , “May be” or “Not”, where Accepted case mean that the proposed model candidate this structure to be the optimal structure to the specific sequence, and “May be “ means that this structure don’t achieve score enough to be Accepted, but also may be selected after some process as the winner one if the model don’t have any Accepted cases to this sequence, and “Not” case main that this structure is rejected , but in our work as will be mentioned in details, we suggest new hypotheses for the “ May be” and “Not” cases to candidate new structures can be the near optimal one.

6. Evaluation rules

In this phase the proposed approach try to set the optimal structure of the specific sequence. This process occurs by looking at the result of inference, were the result can be “Accepted”, “May be” or “Not” cases, first we take only the Accepted cases of the predicted sequence and then evaluate them using first validation criteria “ Accepted case validation “ , which will be explain in the following subsection, and if the predicted sequence fail in obtaining “Accepted” cases, the model will select “May be” cases and apply new validation criteria ” validation B” to set optimal structure, else if the inference produced only “Not” cases, in this case we will suggest new hypothesis that this sequence may be not correct or need to be more accurate, so we backtrack and try to add additional nucleotides to the specific sequence and re-enter the modified sequence to the proposed approach again to set the optimal structure of this sequence. The modification process of the rejected sequence named by “Complete sequence process “, which will explained in individual section.

6.1 “Accepted “case validation

This process don only on the “Accepted” case obtained from inference phase, and as shown in figure 2, if the inference produce only one “Accepted” case, then the proposed approach will select directly this structure to be the optimal structure to the target sequence. And if the inference produce more than one accepted case , we will apply “Accepted case Validation” criteria or rules to select the best one from them.

- Rule 1 : The longest stem of C-rich loop, is the best structure; thus because the binding site in tertiary structure happened between GNRA motif and C-rich stem.
- Rule 2: If the stem of C-rich loop are equal, then look at bottom multi-branch loop if some cases the loop is closed and other is opened we will exclude the structure which contains the opened one.

- Rule 3: Then we suppose that the highest energy structure is the best structure ! yes this not a surprise to us because the special case of IRES (Apical) the best structure should contains some conserved motifs and loops which cause increase in energy.

For example in "AF274010.1" sequence , MFold produce 6 structures for this sequence, and after entering all those different structures to the proposed approach , the result was : 3 May be cases, 1 Not case and 2 Accepted cases (see table 2), so the next step for the model to select all "Accepted" cases which appeared in the result, 2 cases in our example, and then apply "validation A" criteria to chose the best form them, as shown in figure7, both structure have the same value of conserved motifs(3) and multi-branch loops(2) and the same length of C-rich loop stem, the difference rule 2& 3 of "validation A" in closed loop and energy values, so, after applying "validation A", the winner one is the second one although it's have the heights energy, because the previous reason mentioned above and also because its bottom multi-branch loop in closed but in the first structure in opened (see figure 5), result of "Accepted case Validation " explained on table 3.

Table 2: sample from applying inference on apical sequence .

Sequence	#	Energy	Multi-branch loop	Conserved motifs	Case
AF274010.1		-52.10	1	3	**May be**
		low	medium	high	
		-51. 00	1	2	**May be**
		low	medium	medium	
		-50.30	1	3	**May be**
		medium	medium	high	
		-50.0	2	3	**Accepted**
		medium	high	high	
		-50.0	1	2	**Not**
		medium	medium	medium	
		-49.90	2	3	**Accepted**
		medium	high	high	

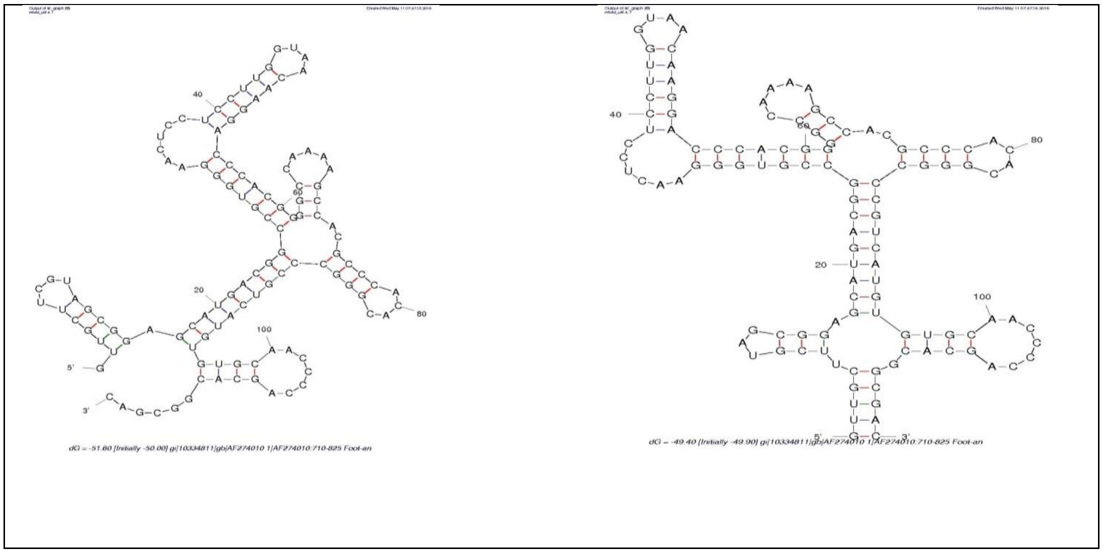
40
20
80
100
5'
3'
dG = -51.60 [Initially -50.00] gi|10334811|gb|AF274010 1|AF274010:710-825 Foot-an
40
80
20
100
5'
3'
dG = -49.40 [Initially -49.90] gi|10334811|gb|AF274010 1|AF274010:710-825 Foot-an

Figure 5: the two Accepted cases of "AF274010.1" produced from proposed approach

Table 3: Accepted case Validation : result

Structure	C-rich stem	Closed loop	energy	result
1	true	false	-50.0	X
2	true	true	-49.90	√

6.2 “May be” case validation

If the proposed approach have no Accepted cases of the sequence, we will collect “May be” cases , and here we have two cases : first : if the sequence length less than reference length (116 nucleotides), then go to complete sequence phase and re-enter the modified sequence to the model(“*second iteration”*). second : if the sequence length is greater than or equal to the reference sequence (>=116), apply “*May be case validation*” : here the lowest energy structure from the candidate structures(May be cases) is the nearest one to the optimal structure.

For example in “AY593756.1” sequence , MFold produce 2 structures for this sequence, and after entering all those different structures in the proposed approach , the result was : 1 May be cases, 1 Not case and 0 Accepted cases (see Appendix A, row number 25), so we will take “May be “ structure and apply “May be case validation” on it : first we check is this sequence is complete length (=> 116), we found that its length = 110, so we jump to complete sequence process which will discussed later, second : after sequence modification (116 nucleotides) we re-entered the modified sequence to the proposed approach again and collect the “Accepted” cases if exist, in our example we obtained 5 structure to the modified sequence ; 2 Accepted cases, 2 May be cases and 1 Not case. As we explained above we will collect all Accepted cases and apply “Accepted case Validation” to select the optimal one from them.

6.3 “Not “ case validation

If we have only “Not” cases; as shown in figure 2, we have two cases, first: if the sequence is less than reference sequence (<116) then go to “complete sequence “ process and re-enter the modified sequence to the proposed approach, second: if the sequence is greater than or equal to the reference sequence (>=116) then the proposed approach take a decision that : this sequence is rejected, or in new word: the proposed approach fail in produce god structure to this sequence. But as discussed in “Result” section: this cases not happened and the proposed approach after the second iteration achieve 100% accuracy.

7. Complete sequence

As shown on figure 2 : If we have only “May be” or “Not” cases, we check sequence length and complete it until we reach to 116 nt length, because the reference sequence (GI: 6318187) has 116 nucleotides length, this phase passed through several steps, fist: we take the rejected sequence (not have any Accepted or May be cases), and type the accession number of it on NCBI portal, then modify query by adding the shortage in the length in 3’ end direction (in the tail), for example : If the sequence of Apical region start from position 765 and end at 874 position, this mean that the length of this sequence is 110, so we will modify query by make the

Apical region start at 765 also but end at 880 to increase length 6 new nucleotides. After obtaining the modified sequence we re-enter it to the proposed approach to pass the model process again to produce Accepted or May be cases this time.

8. Results

After applying proposed approach on 50 apical sequences, as shown in table 4; the proposed approach success in the first iteration to select the suitable structure with 68%(34/50) and after the second round the result raise to be 100%. The data set used is discussed in the net section.

Table 4: results of applying proposed approach on 50 sequences.

#	Sequence	Iteration 1	Iteration 2
1	AF274010.1	√	
2	AJ133357.1	√	
3	AJ133358.1	√	
4	AJ133359.1	√	
5	AM409190.1	√	
6	AM409325.1	√	
7	AY593804.1	√	
8	AY593805.1	√	
9	AY593806.1	√	
10	AY593808.1	√	
11	DQ409183.1	√	
12	DQ409184.1	√	
13	DQ409185.1	√	
14	DQ409186.1	√	
15	DQ409187.1	√	
16	DQ409188.1	√	
17	DQ409189.1	√	
18	DQ409190.1	√	
19	DQ409191.1	√	
20	FJ824812.1	√	
21	AY593751.1	-	√
22	AY593752.1	-	√
23	AY593753.1	√	
24	AY593754.1	√	
25	AY593756.1		√
26	AY593757.1	√	
27	AY593758.1	√	
28	AY593760.1	√	
29	AY593762.1	-	√
30	AY593763.1	-	√
31	AY593764.1	-	√
32	AY593767.1	√	
33	AY593769.1		√
34	AY593771.1	√	

35	AY593774.1		√
36	AY593777.1		√
37	AY593778.1	√	
38	AY593779.1		√
39	AY593780.1	√	
40	AY593781.1	√	
41	AY593783.1		√
42	AY593784.1	-	√
43	AY593785.1	-	√
44	AY593786.1	-	√
45	AY593787.1	√	
46	AY593788.1	√	
47	AY593789.1		√
48	AY593790.1		√
49	AY593792.1	√	
50	AY593794.1	√	

9. Apical data set

From IRES data base [14] we search for FMDV IRES region and found only one strain published on this specialized data base; Foot-and-mouth disease virus (FMDV) strain C, isolate c-s8c1, genomic RNA. We used its GI: 6318187 (Gene bank Id) on NCBI portal [15] and with referring to previous researches which conducted on IRES region to determine start and end positions of domain 3 and its important function region Apical [2,16,17], we take the tested Apical region and make similarity using Blast tool [18] to create data set to Apicals of several FMDV serotypes. The created Apical data set is described at table 5.

Table 5: Apical data set

SeroType	Number of sequences
C	20
A	46
O	19
Asia1	10
Sat 1	0
Sat 2	0
Sat 3	0
Unknown	5
Total	100

10. Contribution

We produced a new approach to select the optimal structure for Apical sequence which saving a lot of effort, time and money for the biologist, where without this work the biologist will test all candidate structures to know the correct one and then try to design the suitable drug to it. And also we produced a new data set for Apical region in FMDV which help other researchers on this point of research .

References

[1] Graham J. Belsham, "*Divergent picornavirus IRES elements*" , Elsevier: Virus Research 139 (2009) 183–192.

[2] Segun Jung and Tamar Schlick, "*Candidate RNA structures for domain 3 of the foot-and-mouth-disease virus internal ribosome entry site*", Nucleic Acids Research, 2013, Vol. 41, No. 3 1483–1495,doi:10.1093/nar/gks1302.

[3] S Lopez de Quinto, Elafuente and E Martenz-Salas, et al., "*IRES interaction with translation initiation factors: Functional characterization of novel RNA contacts with eIF3, eIF4B, and eIF4GII*", RNA (2001) 7:1213-1226, DOI: 10+1017+S1355838201010433.

[4] Amr Badr, Ahmed ElSadek and Alaa.Yassin: "*Computational Based Analysis for Internal Ribosome Entry Site (IRES) and Viral Replication in FMDV*", International Journal of Computer Science Issues (IJCSI), Volume 12, Issue 4, July 2015.

[5] OLGA FERNANDEZ-MIRAGALL and ENCARNACION MARTINEZ-SALAS, "*Structural organization of viral IRES depends on the integrity of the GNRA motif*", rna journal, doi:10.1261/rna.5950603, RNA (2003) 9:1333–1344.

[6] C. Carrillo, E. R. Tulman, G. Delhon, Z. Lu, A. Carreno,Vagnozzi, G. F. Kutish and D. L. Rock, "*Comparative genomics of foot and mouth diseases viruses*", J. Virol. 2005, 79(10):6487. DOI:10.1128/JVI.79.10.6487-6504.

[7] Guohin Lin, Zhipeng Cai, Junfeng Wu, Xin-Feng Wan, Lizhe Xu and Randy Goebel, "*Identifying a few foot-and-mouth disease virus signature nucleotide strings for computational genotyping*", BMC Bioinformatics 2008, 9:279 doi:10.1186/1471-2105-9-279.

[8] Zuker M, Stiegler P. "*Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information".* Nucleic Acids Research. 1981;9(1):133-148.

[9] Mathews DH1, Sabina J, Zuker M and Turner DH. "*Expanded Sequence Dependence of Thermodynamic Parameters Improves Prediction of RNA Secondary Structure*", J Mol Biol. 1999 May 21;288(5):911-40.

[10] http://unafold.rna.albany.edu/?q=mfold/RNA-Folding-Form

[11] Wu TY, Hsieh CC, Hong JJ, Chen CY, Tsai YS," *IRSS: a web-based tool for automatic layout and analysis of IRES secondary structure prediction and searching system in silico*", BMC Bioinformatics. 2009 May 27;10:160. doi: 10.1186/1471-2105-10-160.

[12] Hong JJ, Wu TY, Chang TY, Chen CY," *Viral IRES prediction system - a web server for prediction of the IRES secondary structure in silico.*", PLoS One. 2013 Nov 5;8(11):e79288. doi: 10.1371/journal. Hong JJ1, Wu TY, Chang TY, Chen CY.pone.0079288. eCollection 2013.

[13] http://rna.tbi.univie.ac.at/cgi-bin/RNAfold.cgi

[14] http://iresite.org/

[15] http://www.ncbi.nlm.nih.gov/

[16] FERNÁNDEZ-MIRAGALL, OLGA et al. "*Evidence of Reciprocal Tertiary Interactions between Conserved Motifs Involved in Organizing RNA Structure Essential for Internal Initiation of Translation.*" RNA 12.2 (2006): 223–234. PMC.

[17] Noemí Fernández, Olga Fernandez-Miragall, Jorge Ramajo, Ana García-Sacristán, Nicolás Bellora, Eduardo Eyras, Carlos Briones, Encarnación Martínez-Salas, "*Structural basis for the biological relevance of the invariant apical stem in IRES-mediated translation*", Nucleic Acids Res 2011 Oct 8;39(19):8572-85. Epub 2011 Jul 8.

[18] http://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastn

[19] Du,Z., Ulyanov,N.B., Yu,J., Andino,R. and James,T.L. (2004) " *NMR structures of loop B RNAs from the stem-loop IV domain of the enterovirus internal ribosome entry site: a single C to U substitution drastically changes the shape and flexibility of RNA*" . Biochemistry, 43, 5757–5771.

[20] Lopez de Quinto,S. and Martinez-Salas,E. "*Conserved structural motifs located in distal loops of aphthovirus internal ribosome entry site domain 3 are required for internal initiation of translation*", J. Virol., 71, 4171–4175. (1997).

[21] Robertson,M.E., Seamons,R.A. and Belsham,G.J.,"*A selection system for functional internal ribosome entry site (IRES) elements: analysis of the requirement for a conserved GNRA tetraloop in the encephalomyocarditis virus IRES*", RNA, 5,1167–1179. (1999).