DUJASE Vol. 3(1) 113-122, 2015 (January)

SKnot: A Novel Heuristic Algorithm for RNA Secondary Structure Prediction Including Pseudoknots

Shantanu Saha[1] and Saifuddin Md. Tareeq[1]

[1]*Department of Computer Science and Engineering, University of Dhaka, Dhaka 1000, Bangladesh*

E-mail: shantanucse18@gmail.com and smtareeq@cse.univdhaka.edu

Received on 26. 06. 2014. Accepted for publication on 14.07. 2014.

Abstract

Dynamic programming algorithm for RNA secondary structure prediction including pseudoknot is efficient but expensive. On the other hand heuristic algorithm provide a good alternative for effective solution. We present SKnot, a heuristic algorithm for RNA secondary structure prediction including pseudoknot. Main idea of the algorithm is to find the promising candidate stems which can generate minimum free energy structure including pseudoknot. The algorithm is evaluated on 44 RNA sequences of various types. Experimental result suggest that SKnot predict secondary structure better in most cases in terms of sensitivity and specificity compared to other well known algorithms like PKnotRG, NUPACK and DotKnot.

Keywords: RNA Pseudoknots, Heuristic Algorithm, Dynamic Programming, RNA Secondary Structure.

1. Introduction

RNA carries genetic information for a cell which will express for protein generation. The key factor of RNA is it's 3D structure. This 3D structure can be represented as 2D secondary structure which contains the collection of hydrogen bonds between the base pairs. Figure 1 represents the structure of Hepatitis Delta Virus (HDV) ribozyme [1] sequence. Secondary structure contains stem and loop which has a recursive relation between them. Figure 1 shows stem-base by line. Pseudoknots are formed between unpaired base in loops and the crossed arcs represent the pseudoknots. Pseudoknots also plays important role in protein function. For example, in ribosomal frame-shifting [2] and regulation of translation and splicing [3].

We proposed a new algorithm to generate RNA secondary structure including pseudoknot. Experimental results suggest that our algorithm provide better result in terms of speed, sensitivity and specificity than most of the available algorithms including Dotknot. Time complexity of our algorithm is $O(K^2)$ and space complexity is $O(2*K)$ where K is number of filtered stem. Overall time complexity of RNA secondary structure prediction is $O(K^2*N^3)$ where N is the RNA sequence length.

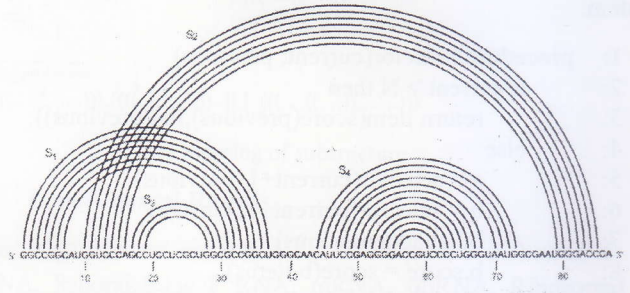


Fig. 1: Hepatitis Delta Virus (HDV) ribozyme structure

2. Background and Related Work

Several researchers proposed dynamic programming algorithms that find the minimum free energy structure from a restricted class that includes certain pseudoknotted structures [4, 5, 6, 7]. Rivas and Eddy has proposed complete recursive algorithm to predict secondary structure. Their algorithm calculate all possible secondary structure to choose best structure including pseudoknots. Rivas and Eddy provides complete model with parameters to calculate the free energy of the structure. However runtime complexity of $O(n^6)$ makes it difficult to run this algorithm for sequence with more than 150 nucleotides (*nt*). Another limitation of this algorithm is that free energy estimates of pseudoknotted structures component used in the algorithm are not optimized. As a consequence, the minimum free energy prediction is often not correct [8]. The pknotsRG-MFE [9] is another dynamic programming algorithm proposed by Reeder and Giegerich with "canonization rules". They proposed three rules for structure generation that helps to reduce the runtime in $O(n^4)$. Their algorithm also provides suboptimal structures [9] and base-pairing probabilities [7].

In contrast heuristic approaches provide no guarantees to find the minimal energy structure but we can calculate RNA secondary structure for larger sequence and they are less restricted than the dynamic programming algorithms. Heuristic algorithms are not limited to sampling from a restricted sub-class, a feature that becomes more important for longer sequence. In last few years, there have been significant advances in the development of heuristic algorithms, leading to improvements in solving RNA secondary structure prediction problems [10]. A disadvantage of this type of approach is that it is not possible to remove or add stems later. Van Batenburg et al. [11] showed that genetic algorithm approach could be promising to predict the pseudoknotted structures by addressing the shortcomings of heuristics algorithm. He described results on a computer simulation of RNA folding pathways using a genetic algorithm for structure prediction

[12]. Another effective algorithm called STAR [12], maintains a list of stems that can be added to a partially formed structure with probability that depends on the free energy of the stem as well as on the free energy of the loop. Their algorithm also includes a mechanism to remove stems and a crossover mechanism for producing new structures from two structures. Their algorithm can correctly predict base pairs ranged from 62% to 87%. One drawback of STAR algorithm is that it requires a user with knowledge to take decision in several steps. Ruan et al. [13] presented a heuristic algorithm for pseudoknots structure prediction called iterative loop matchig (ILM). ILM [13] generates pseudoknotted secondary structures from multiple homologous sequences. A computer simulation of the folding dynamics of an RNA molecule was proposed by Isambert and Siggia [14]. Their method can provide the identification of kinetically trapped states that may be on the folding pathway of the RNA molecule.

3. Our Proposed Algorithm

We have proposed an algorithm as a solution to the RNA secondary structure prediction named SKnot. Existing algorithms are used for local alignment and filtering while a stem selection algorithm is proposed. The pseudocode of the algorithm is given in the SKnotted structure procedure and detail solution is described in the following subsections.

```
1: procedure Sknotted structure
2: Generate stems using Local Alignment Algorithm
3: Filter stems to reduce search space
4: Select best combination of stems:
5: generate secondary structure
6: calculate score for structure and choose MFE structure
7: end procedure
```

3.1 Generating initial list and filtering

GUUGLE [15] is used as local alignment algorithm. GUUGLE can generate exact match between the target sequence and query sequence. GUUGLE returns a fragment of sequence containing *(i,j,l)* where *l* denotes the length of match, *i* denotes the starting position of match and *j* denotes the ending position of match. GUUGLE is used to generate exact match fragment with minimum length 2. GUUGLE assume two *(nt)* is matched, if it is watson-crick base pairs (A-U,C-G) or wobble base pairs(G-U).

Because of that bulge structure is generated from list of stems generated by local alignment with similar technique to Knotseeker [16]. Overlapped bulges are also generated. Figure 2 shows the combination for non overlapping bulge structure generation.

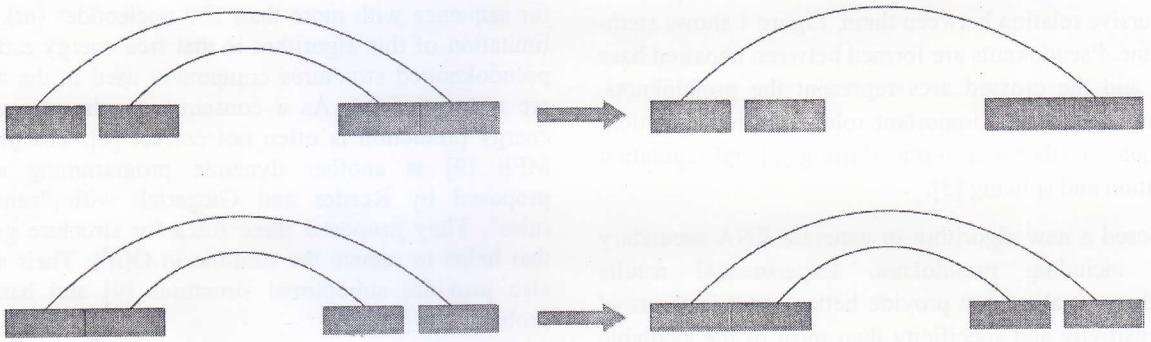


Fig. 2: Construction of bulge loop of size one.

Filtering procedure is based on three restrictions as proposed in [17]. First restriction is that the length of unpaired nucleotide must be three. Second restriction is on the energy of the stems. All the stem with energy <= 2.80 kcal/mol is chosen by extensive experimentation with test cases and the energy is calculated using RNAeval as given by Turner energy parameter [17].

3.2 Selecting Stems and Generating Structure

This section of the paper represents our major contribution. Our task is to choose a single combination of filtered stems generated by GUUGLE [15] which can generate MFE structure. This subproblem is defined as, "From *N* stems in a list choose *k* stems (where $K \leq N$) such that one stem appear only once in any order for which score is minimum". A dynamic programming technique is used to find out the best combination in $O(N^2)$ time. Using the recursive Selector procedure given below our goal is achieved where base case of the procedure is the structure with no stems or only single stem.

```
1: procedure Selector(current; previous)
2:     if current > N then
3:         return item(score(previous),list(previous));
4:     else
5:         a = selector(current+1, previous)
6:         b = selector(current+1, current)
7:         b.list.add(previous)
8:         b.score = score(b.items)
9:         if a < b then
```

```
10:             return a
11:         else
12:             return b
13:         end if
14:     end if
15: end procedure
```

For RNA secondary structure problem all the combination of stems are not valid. If two stem overlap in paired position then they can't be merged. As an example [i:20,j:50,l:10] and [i:28,j:70,l:8] are not valid combination. They can't be merged because they overlap in 28 to 30 index. In this situation two lists are chosen; one list generated with current stem combined with other non-conflicting stems and second list generated without current stem. A list with best MFE structure among them is kept. By removing tail recursion the memory complexity is reduced from to $O(K^2)$ to $O(K)$.

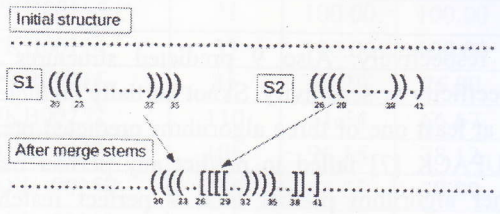
3.3 Proof of Algorithm

There are two subproblems in each steps; 1. Optimal result excluding the current item. 2. Optimal result including the current item. If we can solve these two subproblems correctly then the solution of current step is also correct.

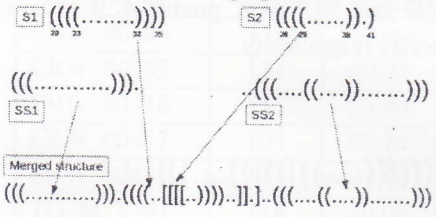
1. **Recurrence:**selector(current,previous)= optimal(selector(current+1,previous),selector(current+1,current) + score(previous,current));
2. **Proof of the correctness of recurrence:** Assume there are one item (a) in list and initial previous item is zero, denotes no item. Then we have two option to choose the optimal result. **1.** Choose the item (a) or **2.** Don't choose the item. If there are two item (a,b) then in the first step we can calculate the optimal value without choosing any item, otherwise we can choose the item set generated by recursion, such as {a}, {b}, {a, b}. There are no other options left to choose the combination of items. This means, proposed recursion can generate all possible combination of item set. Again, the solution technique of this recursion is bottom-up and in each steps, two subproblems are correct so the solution calculated from two subproblem is also correct.
3. **Base cases:** When the current item reached at the end of list, in other word when (current == N) then algorithm return score(previous).

3.4 Scoring Function

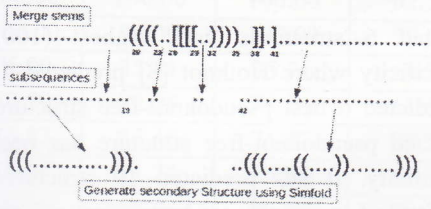
Scoring function receives combination of stems selected by selector procedure. These combinations are merged because they might generate pseudoknot. To calculate secondary structure without pseudoknot in the rest of the section subsequences are generated by removing pseuknot generating (all-paired) nucleotide from RNA sequence as done in [8]. From these subsequences pseudoknot-free secondary structure is generated by using dynamic programming algorithm as done in Simfold [18, 19]. Then pseudoknotted structure is merged with pseuknot-free structure to generated the all combined secondary structure. Scoring function returns the one with minimum free energy structure from pseudoknotted and all combined secondary structure by using thermodynamic parameters [20] together with those Cao and Chen models [21]. These steps are shown in figure 3.



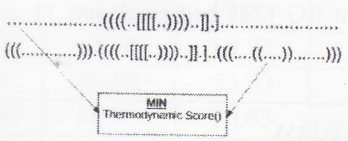
(a) Merging of two selected stems



(c) Merging of subsructure



(b) Generating structure without pseudoknot



(d) Selecting MFE from (a) and (c)

4. Experimental Results

Our algorithm is evaluated on 44 sequence including T-RNA, Ribonuclease P RNA, mRNA, tmRNA, Ribosomal RNA, HIV-1-RT ligand RNA, Viral ribosomal RNA frame-shifting signals, Anti-genomic HDV, Viral RNA, Virus and telomerase RNA. Among them selected RNA sequence 14

are pseudoknot free and other 30 sequence has pseudoknot in structure. The sequence type is summarized in Table 1. In selected sequences 11 are long (130 nt to 354 nt) and other 33 sequences are relatively short (28 nt to 110 nt). To compare the prediction efficiency with other algorithms we computed sensitivity and specificity of the predicted secondary structure is calculated. Table 2 shows the comparative study of sensitivity (SE) and specificity (SP) with different algorithms including ours. Sensitivity and specificity are defined as:

$$SE = \frac{10 \times TP}{TP + FP}$$

$$SP = \frac{100 \times TP}{TP + FN}$$

True positive (TP) denotes the number of correctly predicted base pairs in the predicted structure. False negative(FN) denotes the number of base pairs in original structure that are not predicted in generated structure. Finally False positive(FP) denotes the number of base pairs that are incorrectly predicted in generated structure. Perfectly predicted structure's sensitivity and specificity is 100. The specificity and sensitivity for structure generated by other algorithm is also computed. PknotsRG-mfe 1.3 [22], NUPACK 3.0.4 [7], HotKnots 2.0 [8] and Dotknot 1.3.1 [23] are sued to compare the performance.

Table 1: RNA sequences, used in our experiment.

Sequence ID	Sequence Type
DA0260, DA1280, DD0260, DY4441	T RNA (pseudoknot free)
ASE 00024,ASE 00159,ASE 00161	Ribonuclease P RNA (pesudoknot free)
T4 gene32	mRNA
TMR 00027, TMR 00049, TMR 00007	tmRNA
CRW 00284, CRW 00285,CRW 00418, CRW 00447,CRW 00451	ribosomal RNA (peseudoknot free)
HIVRT32, HIVRT33	HIV1RT ligand RNA
MMTVvpk, T2gene32, BWYV, pKAA	viral ribosomal RNA frame-shifting signals
HDV anti	antigenomic HDV
TYMV, TMV.L, TMV.R CSFV IRES, BVDV IRES	Viral RNA
PKB00001, PKB00045, PKB00038, PKB00137,PKB00168, PKB00016, PKB00256, PKB00143,PKB00155, PKB00114, PKB00216, PKB00252, tobaccomosaicvirus	Virus

SKnot has predicted 6 structures with highest (100) sensitivity and specificity where Hotknot [8] predicted 4. SKnot has also predicted 6 best pseudoknot-free structure out of 14. 3 predicted pseudoknot-free structure has best sensitivity or specificity. It has predicted 19 structures where both sensitivity and specificity are better than at least one of the three other algorithms [22, 7, 8]. On the other hand Hotknot [8], Pknot-RG [22] had predicted 11 and 7 structures respectively. Also 9 predicted structures has highest specificity or sensitivity. Sknot partially predicted 3 for which at least one of three algorithms predicted perfect match. NUPACK [7] failed to predict any perfect match where other algorithm predict total 9 perfect match. 3 predicted structure has best both specificity and sensitivity. The Hknot, Pknot-RG and NUPACK predict 4, 3, 0 best pseudoknot-free structure respectively.

Fig. 4: Dotknot predicted pseudoknot all base pair between 23 and 45 are sued to calculate sensitivity and specificity.

Table 2: Sensitivity and Specificity of the prediction

		Sknot		Hotknot		PKnotsRG		NUPACK	
ID	Length	SE	SP	SE	SP	SE	SP	SE	SP
BWYV	28	55.00	100.00	88.89	100.00	100.00	100.00	55.56	100.00
DA1280	73	100.00	91.30	100.00	91.30	100.00	95.45	100.00	91.30
DD0260	76	90.48	95.00	52.38	57.89	28.57	28.57	33.33	29.17
DY4441	73	100.00	95.45	76.19	69.57	19.05	16.67	71.43	65.22
HDV anti	91	16.67	14.29	16.67	14.29	16.67	14.28	16.67	14.29
HIVRT32	35	41.94	47.27	54.55	100.00	45.45	62.50	45.45	62.5
DA0260	75	27.27	31.58	0	0	77.27	85.00	0	0
CSFV IRES	76	68.00	89.47	28.00	36.84	72.00	85.71	72.00	85.71
BVDV IRES	73	52.00	61.90	28.00	33.33	72.00	90.00	52.00	65.00
TMV.R	105	29.41	31.25	52.94	64.29	67.64	74.19	64.71	70.97
TMR 00007	181	45.45	62.50	54.84	64.15	53.23	58.93	24.19	24.19
MMTV vpk	34	100.00	91.67	100.00	91.67	100.00	91.67	45.45	100.00
pKA A	36	100.00	92.31	100.00	92.31	100.00	92.31	0	0
T2 gene32	33	100.00	100.00	100.00	100.00	100.00	100.00	58.33	70.00
T4 gene32	28	100.00	100.00	100.00	100.00	100.00	100.00	63.64	100.00
HIVRT33	35	45.45	100.00	90.91	100.00	100.00	100.00	0	0
TMV.L	84	68.00	89.47	68.00	85.00	80.00	83.33	52.00	61.90
TYMV	86	84.00	87.50	48.00	60.00	76.00	79.17	68.00	77.27
PKB00001	47	100.00	100.00	66.67	100.00	100.00	100.00	66.67	100.00
PKB00045	41	100.00	100.00	0	0	100.00	100.00	60.00	66.67
PKB00038	41	62.50	45.45	62.50	41.67	0	0	37.50	30.00
PKB00137	133	61.36	65.85	72.73	76.19	86.37	88.37	86.37	88.37
PKB00168	105	73.53	86.21	73.53	86.21	76.47	89.66	82.35	77.78
PKB00016	42	100.00	69.23	100.00	69.23	100.00	69.23	66.67	60.00
PKB00143	71	91.67	88.00	91.67	88.00	75.00	72.00	66.67	84.21
PKB00256	56	55.56	58.82	100.00	100.00	100.00	90.00	55.56	66.67
PKB00155	21	100.00	100.00	100.00	100.00	100.00	100.00	62.50	100.00
PKB00114	33	100.00	100.00	90.00	100.00	90.00	100.00	50.00	83.33
PKB00216	45	64.29	75.00	64.29	75.00	35.71	35.71	64.29	100.00
PKB00252	110	61.54	66.67	61.54	70.59	82.05	84.21	61.54	68.57
ASE_00024	106	96.15	78.12	30.77	25.87	88.46	76.67	92.31	77.42
ASE_00159	186	45.65	38.89	65.22	53.57	71.74	55.00	65.22	50.85
ASE_00161	110	95.65	88.00	95.65	88.00	78.26	62.07	78.26	69.23
TMR_00049	139	40.62	41.94	56.25	64.29	59.38	50.00	53.12	58.62
CRW_00284	132	72.09	70.45	72.09	70.45	67.44	72.05	46.51	52.63
CRW_00285	131	69.77	69.77	67.44	69.05	69.77	69.77	25.58	27.50
CRW_00418	113	95.00	70.37	75.00	75.00	80.00	53.33	75.00	65.22
CRW_00447	104	95.24	100.00	95.24	100.00	80.95	70.83	76.19	100.00
CRW_00451	113	91.30	77.78	91.30	77.78	69.56	55.17	39.13	34.62
TO_M_VIRUS	214	40.00	51.85	58.57	70.69	60.00	66.67	57.14	52.83
TMR_00027	174	61.22	60.00	69.38	73.91	79.59	73.58	57.14	52.83
Telo.human	211	64.00	46.38	60.00	46.15	54.00	42.86	54.00	44.26
CRW_00020	354	60.58	56.70	90.38	81.74	85.57	80.90	62.50	56.52
CRW00054	350	87.25	83.96	82.35	80.77	87.27	85.58	71.57	70.88

Table 3: Sensitivity and Specificity and Ratio comparison between sknot and Dotknot

		Sknot			Dotknot		
ID	Length	SE	SP	R	SE	SP	R
BWYV	28	55.00	100.00	1/1	100.00	100.00	1/1
DA1280	73	100.00	91.30	0/0	100.00	37.04	0/2
DD0260	76	90.48	95.00	0/0	71.43	52.63	0/2
DY4441	73	100.00	95.45	0/0	44.44	23.53	0/2
HDV anti	91	16.67	14.29	1/0	100.00	92.31	1/1
HIVRT32	35	41.94	47.27	1/1	100.00	100.00	1/1
DA0260	75	27.27	31.58	0/2	100.00	54.55	0/1
CSFV IRES	76	68.00	89.47	1/0	68.00	100.00	1/1
BVDV IRES	73	52.00	61.90	1/0	N/A	N/A	1/0
TMV.R	105	29.41	31.25	2/1	93.33	90.32	2/2
TMR 00007	181	45.45	62.50	3/2	74.47	77.78	3/3
MMTV vpk	34	100.00	91.67	1/1	100.0	91.67	1/1
pKA A	36	100.00	92.31	1/1	100.0	91.67	1/1
T2 gene32	33	100.00	100.00	1/1	100.00	100.00	1/1
T4 gene32	28	100.00	100.00	1/1	100.00	100.00	1/1
HIVRT33	35	45.45	100.00	1/0	100.00	100.00	1/1
TMV.L	84	68.00	89.47	3/1	96.00	92.31	3/3
TYMV	86	84.00	87.50	1/1	75.00	66.67	1/1
PKB00001	47	100.00	100.00	1/1	100.00	66.67	1/1
PKB00045	41	100.00	100.00	1/1	0.00	N/A	½
PKB00038	41	62.50	45.45	1/1	0.00	N/A	1/1
PKB00137	133	61.36	65.85	½	100.00	68.42	½
PKB00168	105	73.53	86.21	1/0	0.00	N/A	½
PKB00016	42	100.00	69.23	1/0	N/A	N/A	1/0
PKB00143	71	91.67	88.00	1/1	43.75	47.73	¼
PKB00256	56	55.56	58.82	1/1	N/A	N/A	1/1
PKB00155	21	100.00	100.00	1/1	100.00	100.00	1/1
PKB00114	33	100.00	100.00	1/1	90.00	100.00	1/1
PKB00216	45	64.29	75.00	1/1	80.00	72.73	1/1
PKB00252	110	61.54	66.67	1/1	84.85	93.33	1/1
ASE_00024	106	96.15	78.12	0/1	N/A	0.00	0/1
ASE_00159	186	45.65	38.89	0/2	25.00	15.39	0/1
ASE_00161	110	95.65	88.00	0/0	82.35	50.00	0/1
TMR_00049	139	40.62	41.94	2/2	66.67	61.54	2/2
CRW_00284	132	72.09	70.45	0/0	0.00	0.00	0/2
CRW_00285	131	69.77	69.77	0/0	0.00	0.00	0/5
CRW_00418	113	95.00	70.37	0/0	N/A	N/A	0/0
CRW_00447	104	95.24	100.00	0/0	N/A	N/A	0/0
CRW_00451	113	91.30	77.78	5/3	94.64	94.64	5/5
TO_M_VIRUS	214	40.00	51.85	2/1	0.00	0.00	2/5
TMR_00027	174	61.22	60.00	0/0	66.67	54.55	0/1
Telo.human	211	64.00	46.38	1/0	61.29	51.35	1/1
CRW_00020	354	60.58	56.70	0/2	0.00	0.00	0/1
CRW00054	350	87.25	83.96	0/0	N/A	N/A	0/5

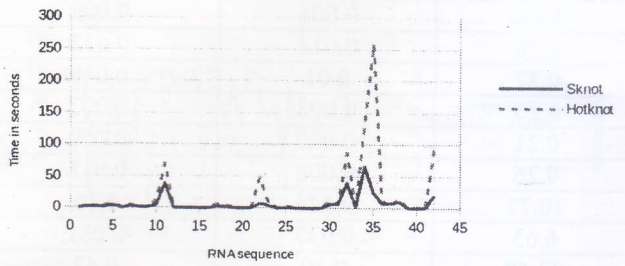


Fig. 5: Runtime(in seconds) comparison between SKnot and Hotknot.

Dotknot [22] predicted only the pseudoknots and the corresponding base pairs position and used these predicted pseudoknotted structure's base pair for sensitivity and specificity calculation. In case of SKnot full structure is used to calculate the sensitivity and specificity. The number of pseudoknot in published structure by the number of pseudoknot in predicted structure is used to calculate the ratio. If (TP + FP = 0) or (TP + FN = 0) or Dotknot predicted no pseudoknot then N/A is used.

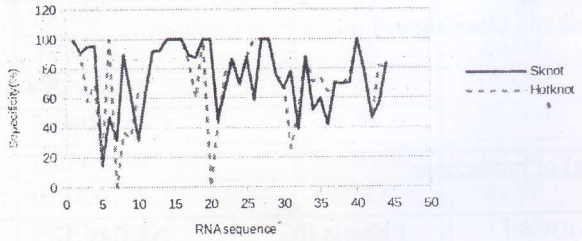
Sknot predicted 6 perfect match out of 9. 19 best structures are predicted where both sensitivity and specificity are higher than dotknot. Also Sknot failed to predict 3 perfect match where Dotknot predicted perfect match for those sequences. Number of perfect match predicted by Dotknot is 6, 10 best structures are predicted over Sknot. Dotknot failed to predict 2 perfect match where Sknot predicted those perfect match successfully. Dotknot failed to predict 12 pseudoknot free structure out of 14 where Sknot predicted 1/2 pseudoknots for 4 pseudoknot free structure out of 14.

Table 4 shows the comparative run time of SKnot, PknotRG and NUPACK algorithms. Run time of PknotsRG is impressive. SKnot run time is low compare to HotKnot. For short sequence the time difference is small but for longer sequence both algorithm's runtime has good difference. Figure 5 shows the run time comparison of SKnot and Hotknot [23] except the last five sequence in the list. These sequences are not included in the graph because the required time for these sequences in much higher and then graph would be illegible for other sequences. Indeed for those five sequences SKnot performs better. Figure 6 shows the sensitivity comparison among SKnot and other algorithm while figure 7 shows the specificity comparison among SKnot and other algorithm.

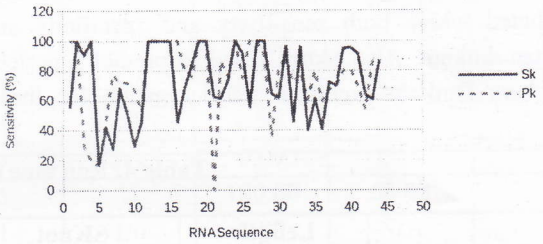
Table 4: Run time (in seconds) of predictions

ID	Length	SKnot	HotKnot	PKnots RG	NUPACT
BWYV	28	0.094	0.227	0.007	0.007
DA1280	73	1.55	2.26	0.015	0.032
DD0260	76	1.49	2.20	0.017	0.037
DY4441	73	1.07	4.57	0.021	0.033
HDV anti	91	3.93	4.13	0.025	0.053
HIVRT32	35	0.02	0.22	0.02	0.02
DA0260	75	2.12	4.43	0.015	0.053
CSFV IRES	76	1.19	1.24	0.016	0.035
BVDV IRES	73	0.89	0.84	0.015	0.032
TMV.R	105	4.25	11.34	0.037	0.082
TMR 00007	181	38.357	69.82	0.25	0.398
MMTV vpk	34	0.029	0.26	0.007	0.009
pKA A	36	0.064	0.284	0.007	0.01
T2 gene32	33	0.027	0.21	0.006	0.008
T4 gene32	28	0.013	0.21	0.007	0.008
HIVRT33	35	0.022	0.263	0.006	0.008
TMV.L	84	1.65	5.69	0.02	0.044
TYMV	86	1.63	4.57	0.02	0.043
PKB00001	47	0.11	0.24	0.007	0.013
PKB00045	41	0.042	0.268	0.007	0.009
PKB00038	41	0.142	0.734	0.007	0.011
PKB00137	133	6.87	47.67	0.10	0.157
PKB00168	105	3.93	8.60	0.038	0.081

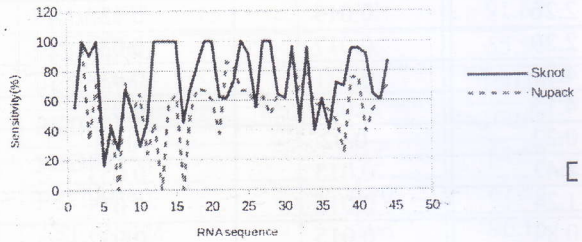
PKB00016	42	0.12	0.362	0.008	0.009
PKB00143	71	0.72	2.38	0.014	0.032
PKB00256	56	0.25	0.37	0.01	0.018
PKB00155	21	0.03	0.23	0.008	0.007
PKB00114	33	0.02	0.21	0.007	0.009
PKB00216	45	0.06	0.26	0.008	0.013
PKB00252	110	5.70	10.77	0.045	0.10
ASE_00024	106	7.98	6.65	0.037	0.081
ASE_00159	186	39.57	87.48	0.30	0.47
ASE_00161	110	2.02	3.84	0.043	0.091
TMR_00049	139	7.73	12.10	0.080	0.159
CRW_00284	132	8.28	4.84	0.102	0.167
CRW_00285	131	10.41	15.15	0.076	0.450
CRW_00418	113	0.803	1.16	0.057	0.086
CRW_00447	104	0.48	1.03	0.031	0.067
CRW_00451	113	1.30	2.59	0.042	0.091
TO_M_VIRUS	214	65.74	1089.92	0.045	0.65
TMR_00027	174	25.15	258.87	0.20	0.322
Telo.human	211	182.718	95.209	0.478	0.064
CRW_00020	354	1146.39	1698.13	3.165	3.104
CRW00054	350	86.57	2428.418	2.803	2.922



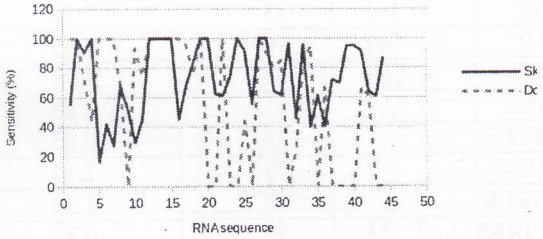
(a) Graph between Sknot and HotKnot



(b) Graph between Sknot and PknotsRG

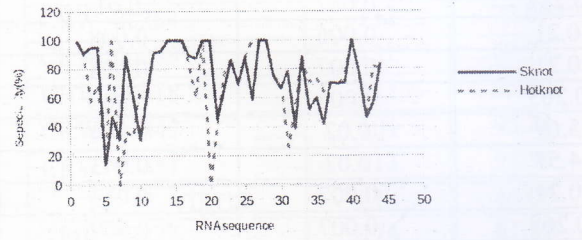


(c) Graph between Sknot and NUPACK

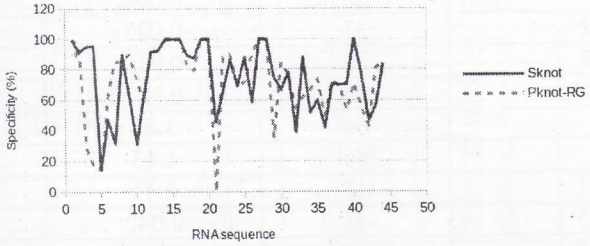


(d) Graph between Sknot and DotKnot

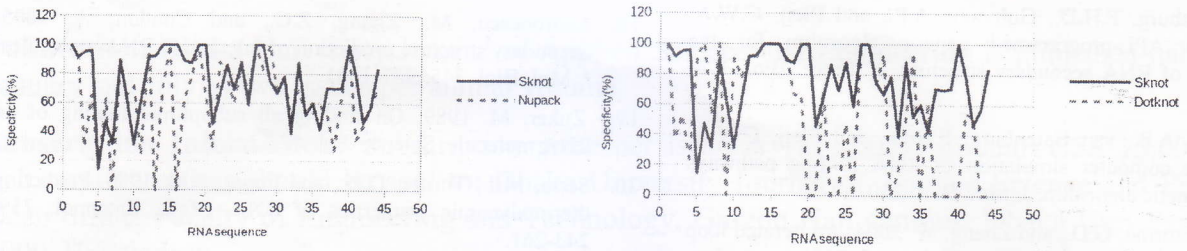
Fig. 6: Comparison of sensitivity(SE) between SKnot and other algorithms.



(a) Graph between Sknot and HotKnot



(a) Graph between Sknot and PknotsRG



(c) Graph between Sknot and NUPACK (d) Graph between Sknot and DotKnot

Fig. 7: Comparison of specificity(SP) between SKnot and other algorithms.

5. Discussion

As heuristic approach is used in our algorithm to reduce the search space it is not guaranteed to find the optimal structure. Sknot predicted 6 perfect match out of 44 sequence where Hotknot can predict 4 prefect match, 7 by PknotRG and zero by NUPACK. For BWYV sequence, we found different optimal structure with respect to energy model with Dirks and Pierce (DP) [7] and Cao and Chen (CC) [21] models. Also the energy of predicted structure of BWYV using CC energy model is low with respect to original or published structure's energy. Original structure's free energy is -2.116 kcal/mol where predicted structure's free energy is -8.05 kcal/mol. Original structure's energy should be lowest but energy model returns lowest energy for other structure, indicating the weakness of energy model.

Sknot predicted 5 structure with best sensitivity and specificity value for long pseudoknot free structure out of 10 and 2 structure has best sensitivity or specificity value. Sknot failed to predict best structure for 1 pseudoknot free sequence. On the other hand PknotRG predicted 2 (CRW_00285,ASE_00159) structure where CRW_00285 has the same sensitivity and specificity predicted by Sknot. Also Sknot predict 2 best structures out of three short pseudoknot free structure.

Sknot has sensitivity < 50.00 on 9 structure out of 44 structure. One of the this 9 structure is the best predicted structure with respect of other algorithms. 19 structure has sensitivity > 80.00. On the other hand Hotknot predict 14 structure which sensitivity > 80.00. Future work will be to increase prediction efficiency and make it faster using better alignment algorithm and energy evaluation technique respectively.

6. Conclusion

There is no doubt that RNA secondary structure prediction with pseudoknot is very important task. Our algorithm can predict H-type pseudoknotted structure as well as pseudoknot free structure more efficiently. The weak point of our algorithm is that it's not fast like PknotRG or NUPACK. There are more options available to improve our algorithm by changing energy model, and faster and efficient local alignment algorithm. In future, we will extend our algorithm to adopt more complex type of pseudoknot.

References

1. Isambert, H. and Siggia, E.D. 2000. Modeling RNA folding paths with pseudoknots: Application to hepatitis delta virus ribozyme. Proc.Natl. Acad. Sci. 97: 6515-6520.

2. Giedroc, D.P., Theimer, C.A., and Nixon, P.L. 2000. Structure, stability and function of RNA pseudoknots involved in stimulating ribosomal frameshifting. J. Mol. Biol. 298: 167-185.

3. Draper, D.E., Gluick, T.C., and Schlax, P.J. 1998. Pseudoknots, RNA folding and translational regulation. In RNA structure and function (eds. R.W. Simons and M. Grunberg-Manago), pp. 415-436. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

4. Rivas, E. and Eddy, S.R. 1999. A dynamic programming algorithm for RNA structure prediction including pseudoknots. J. Mol. Biol.285:2053-2068.

5. Uemura, Y., Hasegawa, A., Kobayashi, S., and Yokomori, T. 1999. Tree adjoining grammars for RNA structure prediction. Theor. Comp. Sci. 210: 277-303.

6. Lyngs, R.B. and Pedersen, C.N. 2000. RNA pseudoknot prediction in energy-based models. J. Comp. Biol. 7(3): 409-427.

7. Dirks, R.M. and Pierce, N.A. 2003. A partition function algorithm for nucleic acid secondary structure including pseudoknots. J. Comput. Chem. 24: 1664-1677.

8. J Ren, B Rastegari, A Condon, HH Hoos, 2005. Hotknot: Heuristic prediction of RNA secondary structures including pseudoknots, RNA. 2005 Oct;11(10):1494-504.

9. Reeder, J. and Giegerich, R. 2004. Design, implementation and evaluation of a practical pseudoknot folding algorithm based on thermo- dynamics. BMC Bioinformatics 5: 104 .

10. Hoos, H.H. and Stu tzle, T. 2004. Stochastic local search: Foundations and applications. Morgan Kaufmann, San Francisco, CA.

11. Van Batenburg, F.H.D., Gultyaev, A.P., and Pleij, C.W.A. 1995. An APL-programmed genetic algorithm for the prediction of RNA secondary structure. J. Theor. Biol. 174: 269-280.

12. Gultyaev, A.P., van Batenburg, F.H.D., and Pleij, C.W.A. 1995. The computer simulation of RNA folding pathways using a genetic algorithm. J. Mol. Biol. 250: 37-51.

13. Ruan, J., Stormo, G.D., and Zhang, W. 2004. An iterated loop matching approach to the prediction of RNA secondary structures with pseudoknots. Bioinformatics 20: 58-66.

14. Isambert, H. and Siggia, E.D. 2000. Modeling RNA folding paths with pseudoknots: Application to hepatitis delta virus ribozyme. Proc. Natl. Acad. Sci. 97: 6515-6520.

15. Gerlach, W. and Giegerich, R. 2006. GUUGle: A utility for fast exact matching under RNA complementary rules including G-U base pairing. Bioinformatics 22: 762-764.

16. J Sperschneider, A Datta - RNA, 2008. KnotSeeker: Heuristic pseudoknot detection in long RNA sequences. RNA. 2008 Apr; 14(4): 630–640.

17. Zuker, M., Mathews, D.H., and Turner, D.H. 1999. Algorithms and thermodynamics for RNA secondary structure prediction: A practical guide. In RNA biochemistry and biotechnology Kluwer Academic Publishers, Dordrecht, The Netherlands.

18. Andronescu, M., Zhang, Z.C., and Condon, A. 2005. Secondary structure prediction of interacting RNA molecules. J. Mol. Biol. 345: 987-1001.

19. Zuker, M. 1989. On finding all suboptimal folding of an RNA molecule. Science 244: 48-52.

20. Serra, M.J., Turner, D.H., and Freier, S.M. 1995. Predicting thermodynamic properties of RNA. Meth. Enzymol. 259: 243-261.

21. Cao, S. and Chen, S.-J. 2005 Predicting RNA folding thermodynamics with a reduced chain representation model RNA 1118841897.

22. Reeder, J. and Giegerich, R. 2004. Design, implementation and evaluation of a practical pseudoknot folding algorithm based on thermo- dynamics. BMC Bioinformatics 5: 104 .

23. Jana Sperschneider and Amitava Datta, DotKnot 2010. Pseudoknot prediction using the probability dot plot under a refined energy model. Nucleic Acids Res. 2010 Apr; 38(7): e103.