On the Effectiveness of Rebuilding RNA Secondary Structures from Sequence Chunks

Michela Taufer[1], Thamar Solorio[2], Abel Licon[1,3], David Mireles[3], and Ming-Ying Leung[4]

[1] *Dept. of Computer & Inf. Sciences*
University of Delaware

[2] *Dept. of Computer of Science*
The University of Texas at Dallas

[3] *Dept. of Computer Science*
The University of Texas at El Paso

[4] *Dept. of Math. Sciences and Bioinformatics Prog.*
The University of Texas at El Paso

mtaufer@acm.org, tsolorio@hlt.utdallas.edu,
{alicon2, dvmireles}@miners.utep.edu, mleung@utep.edu

Abstract

Despite the computing power of emerging technologies, predicting long RNA secondary structures with thermodynamics-based methods is still infeasible, especially if the structures include complex motifs such as pseudoknots.

This paper presents preliminary results on rebuilding RNA secondary structures by an extensive and systematic sampling of nucleotide chunks. The rebuilding approach merges the significant motifs found in the secondary structures of the single chunks. The extensive sampling and prediction of nucleotide chunks are supported by grid technology as part of the RNAVLab functionality. Significant motifs are identified in the chunk secondary structures and merged in a single structure based on their recurrences and other statistical insights. A critical analysis of the strengths, weaknesses, and future developments of our method is presented.

1. Background and Significance

RNA secondary structure prediction can provide insights in the reconstruction of 3D RNA structures and their functionality. Study of RNA secondary structures is a field that is raising the attention of the scientific community; new insights in the field point out the need for supporting experimental research with computational results. The latter can help to narrow down the space of the experiments and therefore the cost to obtain results.

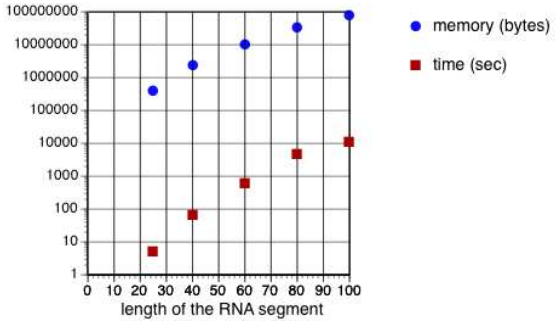

Figure 1. Time and memory usage by Pknots-RE for PseudoBase sequences with different lengths

Despite the computing power of supercomputers and emerging advanced technologies, e.g., multi-core architectures, the prediction of secondary structures for long RNA using thermodynamics-based methods e.g., Zuker and Turner [14], is still infeasible, especially if the structures include complex secondary structures such as pseudoknots. The space and time required for accurate predictions of pseudoknots based on free energy minimization algorithms grow very rapidly with the sequence length. Figure 1 shows the time and memory (in logarithmic scale) allocated for the prediction of RNA pseudoknots with various lengths using one of the most accurate prediction programs, Pknots-RE [11]. The algorithm underlying Pknots-RE has a run time and memory demand in the order of n^6 and n^4 re-

spectively, where n is the length of the input sequence [11]. The program conducts an exhaustive search for the optimal structure with the lowest free energy and has the capability to predict rather complex structures, even some non-planar structures for short RNA segments of up to 200 nucleotides. Some variations of this algorithm have been developed to get around the memory and computing time demands by restricting the types of pseudoknots to be predicted and the RNA sequence lengths to keep computation time and storage size under control. For instance, the program Pknots-RG [10] limits the types of pseudoknots to simple structures for longer segments, up to 800 nucleotides. However, a large variety of pseudoknots occur in reality. Their omission from computational methods might significantly affect the prediction accuracy. Even simplified programs are not able to predict secondary structures on the order of thousands of nucleotides.

We have observed that most pseudoknots experimentally observed are formed by RNA segments whose lengths are less than 200 nucleotides. An analysis of the length of the pseudoknots in PseudoBase [13], which collects 245 of such RNA segments, showed that about 95% of these segments have lengths that range between 20 and 200 nucleotides. Moreover, the range of lengths between 30 (lower quartile) and 67.5 (upper quartile) nucleotides covers 50% of all segments. This observation leads us to the idea of developing a strategy for cutting off the viral genome into segments or chunks of length no more than 200 bases, and distributing the task of structure prediction of each chunk to be done simultaneously on different computers.

Ideally, if two chunks cut from the same RNA sequence overlap each other, the predicted structures on their overlapping part should be consistent with one another. Such consistency is important for the final structure assembly. In our preliminary trials, we have observed that arbitrarily cutting the RNA sequence into overlapping segments is not advantageous for consistency. It is well conceivable that when an arbitrary cut goes through the middle of an inversion, the bases forming the pairings do not get into the same segment causing the omission of the structure on that prediction. For instance, consider a 100 base piece of the Severe Acute Respiratory Syndrome (SARS) coronavirus genome, which is one of the coronavirus genomes analyzed by Chew et al. in [4], from position 25884 to 25983 and another piece from position 25923 to 26022. When the program Pknots-RE is applied to these segments, two predictions are produced which are shown in Figure 2. Note that, over the stretch of 62 bases when the two pieces overlap one another, the two predictions are different. This kind of inconsistency poses a serious problem when the predicted structures of the segments need to be assembled.

This paper addresses this challenge and presents a method for rebuilding RNA secondary structures by systematically sampling nucleotide chunks from a RNA sequence and rebuilding the secondary structure of the longer sequence from the motifs found in the secondary structures of the chunks (i.e., stem-loops and pseudoknots). The extensive and systematic sampling of nucleotide chunks is vital for the success of our method and the inconsistency outlined above; the computing power needed for the prediction of the numerous chunks is provided by grid technology as part of the RNAVLab functionality [12]. Motifs are identified in the chunk secondary structures and merged in a single structure based on their recurrences and other statistical insights. Rigorous statistical analysis are used to identify weaknesses and strengths of the proposed sampling and the rebuilt algorithm.

This paper is organized as follows: Section 2 provides a overview of related work. Section 3 presents our method for sampling nucleotide chunks, predicting and identifying relevant motifs, and rebuilding whole secondary structures from the chunk motifs. Section 4 statistically quantifies the effectiveness of our method. Section 5 summarizes the main contributions and lists future work.

2 Related Work

Most of the previous work on finding consensus motifs takes as input a set of primary sequences and generates as output the set of structural motifs identified, and the differences lie on the search strategy for identifying common motifs. For instance, work presented in [8, 1] uses suffix arrays for efficiently exploring the space of valid secondary structures in their Seed method. In Seed, the search space is constrained by the seed sequence, which is just one of the sequences in the set used to instantiate valid motifs. Seed ranks motifs using a metric that combines the entropy of the segment with the free energy of the secondary structure, as computed by MFOLD [15]. This ranking function yielded good results, the top motifs had also the highest Matthews Correlation Coefficient [7]. A drawback of Seed is the fact that it is limited to find patterns in stem regions only, that is, no loops or pseudoknots can be identified by the Seed method.

Ashlock and Schonfeld propose a depth annotation scheme to identify common motifs that uses an evolutionary algorithm to cluster folds by projecting them in a two dimensional Euclidean space [2]. The intuition behind this approach is that similar folds will be placed closer by the projection algorithm. To identify motifs we need to analyze the output of the projection. Since the method provides a visual representation of the similarity between bricks, it is simple to identify motifs by just looking for clusters. However, as the number of bricks increases, spotting the clusters become less straightforward and we need the help of a clustering algorithm. Another shortcoming of this method

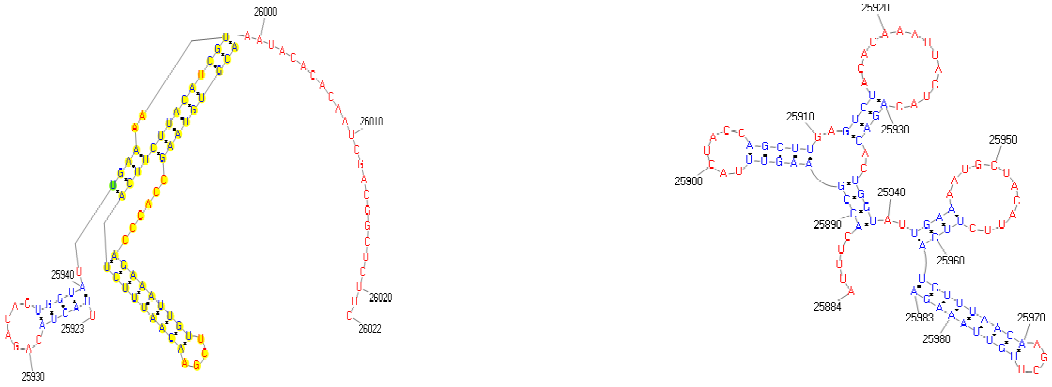

Figure 2. Pknots-RE predictions of SARS segment 25884 base to 25983 base (left) and SARS segment 25923 base to 26022 base (right)

is that the distances between the pairs of depth annotations depend on a specific size of segment. Thus prior knowledge of the sequences is needed in order to define an appropriate window size. This method can identify pseudoknots by assigning a unique identifier to stems.

There are other approaches to motif finding, see for example [3], but most of them give the desired results provided the secondary structure is not complex, that is, no loops or pseudoknots are included, or if we have enough prior knowledge regarding the identity of the motifs. On the contrary, our automated method targets motifs that are as general as possible and exhaustively explores the search space of all the sequences of nucleotides. It is a strictly structural method in the sense that currently we only look at the secondary structure predicted by Pknots-RG –for the experiments presented here. Our preliminary results show that our method can find motifs as simple as small stems and as complex as pseudoknots and loops.

3. Methodology

The method we propose in this paper uses the RNAVLab environment [12] to: (1) sample nucleotide subsequences, or chunks, from larger sequences and predict the secondary structures for each corresponding chunk (Section 3.1); (2) identify common motifs in the partial predictions (Section 3.2), and (3) rebuild the final secondary structure by merging the motifs identified in the previous step (Section 3.3). RNAVLab (RNA Virtual Laboratory) is a unified computational environment for the study of RNA secondary structures that combines sampling of nucleotides sequences, predictions based on different codes and supported by grid computing technology, as well as analysis of large sets of secondary structures.

3.1 Sampling and Predicting Secondary Structures

Our current sampling approach is straightforward. We use sliding windows of nucleotide chunks that progressively grow in length and sliding steps. For each sequence, we systematically generate several sets of chunks that are forwarded to the prediction module in the RNAVLab; each set results in a single rebuilt secondary structure.

Each set of chunks has a fixed size (*window size*) and a fixed sliding step (*window step*). The several chunks in a set are generated by progressively sliding the fixed-size window of a fixed number of steps and each time sampling the nucleotides within the window. The process is repeated to generate the several sets by increasing the window sizes and/or the window steps every time we generate a new set of chunks. Window sizes are always increased by 5 bases. The max length of a window is $n/2$, where n is the length of the RNA sequence we want to rebuild. Window steps range from 1 base to $w - 1$ bases, where w is the window size.

Given a set of chunks, we predict their secondary structures in parallel by using the structure predictor component of RNAVLab. This component harnesses heterogeneous computing resources across the University of Texas at El Paso (UTEP) campus to rebuild RNA secondary structures from RNA segments, using different prediction codes. Currently RNAVLab supports the following prediction codes: Pknots-RE, Pknots-RG [10], and NuPack [9]. In this paper we use the Pknots-RG code but the predictions can be easily extended to the other two codes.

```
Let S be a set of n secondary structure sequences
Let M be the set of all valid structural motifs
1. M ← {}
2. For every s_i in S
   2.1. Generate a new set VS_i with all the valid motifs in s_i
   2.2 M ← {M ∪ VS_i }
3. For every motif m_i in M
   3.1 Search and store the location of all occurrences of m_i in S
4. For every motif m_i in M
   4.1 Rank m_i according to the set of criteria C
4. 2 Remove m_i if the scoring is below a given threshold
```

Figure 3. Pseudocode for the Motif Identifier

3.2 Identification of Common Motifs in Predictions

The identification of common motifs is performed by the Motif Identifier in RNAVLab. The Motif Identifier first identifies all the valid secondary structures, from the most general (i.e., a hairpin comprising a single base pair) to the most complex (i.e., pseudoknots), that can be generated from the input of secondary structures. Then by using an associative array of linked lists, our tool finds and stores the locations of each substructure generated in the previous step. Figure 3 shows the pseudocode of the tool. To narrow down the number of motifs and identify the most relevant ones, ranking techniques are applied. Ranking criteria include: the frequency of the motif over the maximal number of possible occurrences, the number of bonding nucleotides, the length of the secondary structure, and the motif location in the RNA segment. Other possible ranking criteria can include information of the primary structure such as the percentage of bases correctly matched, and/or free energy of the structure. In this paper we score motifs based on their frequency (f), number of base pairs (s), and the length of the overlapping region (o):

$$score = \frac{f * s}{o} \quad (1)$$

The simple intuitive motivation behind this scoring function is that more accurate secondary structures are more likely predicted by longer overlapping structures with higher frequency.

3.3 Rebuilding Secondary Structures from Common Motifs

To rebuild the final secondary structure out of the chunk motifs, we use the scoring function presented previously. We project motifs, in descending order according to their score, into a final structure until there are no more mutually exclusive motifs in the set. In other words, we only project different motifs found in chunks when they do not overlap with each other. As part of the rebuilding algorithm, we also define the minimum frequency that a motif present in overlapping chunks has to meet in order to be projected in the resulting rebuilt sequence (*threshold*). Threshold values can range from 1 to 9. Finally, we compute the energy of the rebuilt structures as a whole by using the same energy algorithm used in Pknots-RG and NUPack.

4 Analysis

In this section we address two important analysis components. First, we quantify the capability of our rebuilding algorithm to capture the secondary structure observed experimentally. We compare performance and accuracy (in terms of sensitivity and selectivity) of our rebuilding algorithm based on nucleotide chunks against a traditonal algorithm using the same prediction code and the entire sequence. Second, we statistically quantify the effectiveness of our naive approach for sampling nucleotide chunks and we measure whether the extensive sampling and predictions can compensate for the fact that no attention is paid to the type of nucleotides in the chunks, i.e., if there are palindrome sequences or not.

4.1 Experiment Set-up

Long RNA secondary structures, i.e., of the order of thousands of nucleotides, that have been experimentally validated are rare. When available, our method can deal with the prediction of these sequences but other methods that predict secondary structures using the entire sequence as a whole cannot, making a comparison between the two approaches infeasible. Therefore, for our analysis in this paper we used the 39 longest nucleotide sequences from Group A in [5] that have lengths ranging from 100 to 482 bases and are still predictable as a whole by the Pknots-RG code. Note that since we are not considering the exact same set as in [5], we cannot perform a direct comparison against those results.

The sampling, motif identification, and rebuilding were executed on the RNAVLab server. Window sizes, window steps, and thresholds used in the experiments are defined in Section 3. The predictions were performed on a 64-node cluster (each node consists of 2 AMD Opteron processors running at 2 GHz with 4 Gigabyte of RAM and a local 120 Gigabyte hard disk) that is part of the on-campus grid resources of RNAVLab. The accuracy of predictions is measured in terms of sensitive (i.e., ability to predict all true pairs) and selectivity (i.e., ability to only predict true pairs).

Predictions are compared with the experimental secondary structures provided in [5].

Table 1 presents the 39 sequences (*Sequence*), their length in bases (*Length*), their number of rebuilt structures including those that, when compared with the experimental secondary structures, have sensitivity and selectivity equal to zero (*Predictions Attempts*), the number of rebuilt structures that have a positive sensitivity and selectivity (*Predictions Used*), the total time in seconds needed for all the chunk predictions on the cluster (*Rebuilt Time*), and the time in seconds used for the prediction of sequences as a whole when using Pknots-RG (*Pred. Time*). The table outlines the high cost in terms of computation needed for our approach. RNAVLab makes our approach feasible by allowing us to perform the computation on idle resources across the campus.

4.2 Accuracy of Rebuilt Structures

In Table 2 we present a summary of the accuracies: the oracle or upper bound on sensitivity and selectivity for our method (*Rebuilt Sen.* and *Rebuilt Sel.*) is compared with the sensitivity and selectivity of Pknots-RG when considering the entire sequence for prediction (*Pred. Sen.* and *Pred. Sel.*) as well as the sensitivity and selectivity achieved by our algorithm when selecting those structures with the lowest free energy (*Min En.*, *Min En. Sen.*, and *Min En. Sel.*).

Since we are using Pknots-RG for the prediction of the chunks, intuitively we would expect our algorithm to achieve results that are at most equally accurate as those achieved by this prediction code when predicting the whole sequence. However, because we are allowing the prediction of chunks starting at different positions in the primary structure, our method can find sequences that are very different from those predicted by the code on the entire sequence. Out of the total 39 sequences presented in the table, the oracle outperformed Pknots-RG on sensitivity and/or selectivity for 24 sequences (see bold values in the table). The selection criteria based on the minimum free energy (*Min En. Sen.* and *Min En. Sel.*) is not as accurate though, only in 7 out of the 39 cases did this criteria yield better or equal results than Pknots-RG.

The results of our rebuilding algorithm are promising, especially considering the potential of our approach in overcoming the limitations of current prediction methods on the length and complexity of the sequences. Currently, the most salient weakness of our method involves the selection of the final rebuilt structure. The minimum free energy is not by itself a good factor for selection, this is probably due to what is already common belief that native structures will often be near-optimal in terms of the minimum free energy. Section 5 discusses in detail our ongoing work to improve this selection criteria.

4.3 Effectiveness of Sampling Approach

In cutting an RNA sequence into chunks of overlapping sequences, we experimented using various window sizes and window step sizes. In rebuilding the overall structures from the chunks, different threshold values have been used. We have noticed that as values of these parameters vary, the overall accuracy, measured by sensitivity and selectivity of the rebuilt structures, also change. In order to check whether any significant systematic relationship exists between the accuracy of the rebuilt structures and the parameters, we carried out a multiple regression analysis on each of the 39 sequences in Table 2 with sensitivity and selectivity as response variables and (window size, window step, threshold) as predictor variables. In all except one sequence, both sensitivity and selectivity are significantly (p-value < 0.005) related to the three predictor variables. Both response variables correlate positively with window size, but negatively with window step and threshold. The positive correlation with window size agrees with our expectation that having a larger sequence chunk, which constitutes a larger portion of the whole RNA molecule, in a single prediction should generally be beneficial to the accuracy of the rebuilt structure. On the other hand, a larger window step would mean that successive sequence chunks overlap less with each other so that it is easier to miss those secondary structures spanning both chunks but not captured within either one, resulting in the negative correlation with the window step parameter. The negative correlation of threshold with structure accuracy implies that every motif detected in a sequence chunk should be taken into account in the rebuilt structure. A very strong positive correlation between sensitivity and selectivity (correlation coefficient > 0.9) has been detected in each of the 39 sequences while the predictor variables are being varied. This suggests that our structure rebuilding approach can be made highly effective simultaneously in both measures of accuracy. It is also interesting to note that the minimum free energy of a rebuilt structure generally shows a negative correlation with sensitivity, but a positive correlation with selectivity, suggesting that the minimum free energy does not necessarily reflect the accuracy of the rebuilt structure. While the minimum free energy is the quantity used pervasively in many secondary structure prediction algorithms for determining what the optimal structure is, there seems to be a necessity for seeking an alternative measure.

4.4 Discussion

The proposed approach can be used for the identification of motifs across prediction codes, given the same segment of nucleotides. This can be useful for validation of prediction techniques across codes with different methodologies

Table 1. Performance comparison of predictions performed with our rebuilding algorithm based on sampled chunks and the same predictions using Pknots-RG and the entire sequence

Sequence	Length	Predictions Attempted	Predictions Used	Rebuilt Time (sec)	Pred. Time (sec)
RF00167_A	100	1701	362	680.40	0.18
RF00374_A	101	1701	351	683.10	0.20
RF00499_A	102	2142	495	831.78	0.21
RF00162_A	103	2142	586	835.47	0.22
RF00198_A	104	2142	451	854.28	0.23
RF00435_A	109	2142	499	905.22	0.26
RF00485_A	114	2628	449	1118.52	0.31
RF00020_A	115	2628	508	1141.65	0.32
RF00001_A	117	2628	678	1168.20	0.23
RF00383_A	117	2628	604	1157.22	0.32
RF00286_A	118	2628	510	1161.36	0.31
RF00463_A	127	3159	824	1490.04	0.42
RF00182_A	129	3159	733	1519.47	0.41
RF00373_A	133	3735	951	1809.54	0.44
RF00290_A	140	3735	974	1924.92	0.53
RF00004_A	145	4356	1277	2348.55	0.61
RF00484_A	149	4356	1246	2376.90	0.66
RF00025_A	152	5022	854	2818.53	0.72
RF00050_A	157	5022	1252	2987.10	0.84
RF00171_A	168	5733	1490	3807.09	1.13
RF00387_A	168	5733	1549	3706.74	1.01
RF00259_A	169	5733	1426	3743.64	1.1
RF00232_A	170	5733	1423	3775.05	1.08
RF00391_A	171	5733	1466	3735.18	0.96
RF00013_A	183	7290	1885	5428.89	1.44
RF00458_A	202	9027	2618	8174.16	2.03
RF00193_A	273	16524	5267	31195.71	6.66
RF00231_A	275	16524	4519	31961.34	6.79
RF00503_A	293	19071	5988	47340.99	9.41
RF00030_A	297	19071	5845	47387.79	9.47
RF00216_A	302	20412	4855	51796.98	9.85
RF00010_A	312	21798	7393	61489.98	10.59
RF00009_A	320	21798	6056	62566.20	11.67
RF00100_A	330	23229	6393	74524.86	13.56
RF00036_A	337	24705	8555	86616.63	14.6
RF00209_A	379	31059	10631	154066.14	22.35
RF00024_A	451	43911	12050	471230.46	53.44
RF00210_A	462	47988	17172	526495.41	50.79
RF00177_A	482	52245	19114	668287.08	59.13

Table 2. Accuracy comparison (in terms of sensitivity and selectivity) of the upper bound rebuilt predictions based on sampled chunks, the same predictions with the entire sequence, and the rebuilt prediction with lowest free energy

Sequence	Length	Rebuilt Sen.	Rebuilt Sel.	Pred. Sen.	Pred. Sel.	Min En.	Min En. Sen.	Min En. Sel.
RF00167_A	100	0.73	0.64	1.00	0.79	-23.50	0.64	0.47
RF00374_A	101	**0.81**	0.65	0.81	0.67	-41.60	**0.81**	0.65
RF00499_A	102	0.76	0.71	0.91	0.86	-34.60	0.73	0.59
RF00162_A	103	0.70	0.52	0.85	0.66	-24.10	0.59	0.46
RF00198_A	104	**0.96**	**0.70**	0.92	0.55	-35.40	**0.92**	**0.63**
RF00435_A	109	0.97	0.97	1.00	1.00	-52.90	0.62	0.46
RF00485_A	114	0.54	0.34	0.71	0.37	-24.20	0.33	0.18
RF00020_A	115	0.73	0.58	0.97	0.74	-36.90	0.73	0.58
RF00001_A	117	0.55	0.41	0.82	0.61	-39.40	0.55	0.40
RF00383_A	117	**0.75**	**0.36**	0.75	0.32	-36.70	0.06	0.02
RF00286_A	118	0.86	0.51	0.95	0.57	-37.50	0.86	0.51
RF00463_A	127	**0.80**	**0.80**	0.61	0.42	-53.10	0.41	0.29
RF00182_A	129	0.53	0.36	0.83	0.56	-35.76	0.50	0.30
RF00373_A	133	**0.64**	**0.45**	0.64	0.18	-22.68	0.18	0.05
RF00290_A	140	**0.93**	**0.93**	0.77	0.49	-33.40	0.73	0.54
RF00004_A	145	**0.97**	**0.81**	0.77	0.48	-45.90	**0.77**	**0.48**
RF00484_A	149	**0.58**	**0.44**	0.39	0.21	-40.70	0.27	0.15
RF00025_A	152	**0.73**	**0.53**	0.70	0.49	-21.56	0.55	0.35
RF00050_A	157	0.56	**0.30**	0.68	0.28	-71.00	0.24	0.09
RF00171_A	168	**0.94**	**0.61**	0.91	0.57	-47.50	**0.91**	**0.60**
RF00387_A	168	0.73	0.71	0.96	0.96	-50.34	0.63	0.53
RF00259_A	169	0.52	0.38	0.59	0.47	-33.00	0.34	0.24
RF00232_A	170	0.61	**0.49**	0.63	0.49	-58.90	0.59	0.45
RF00391_A	171	**0.66**	**0.41**	0.50	0.27	-45.30	0.28	0.13
RF00013_A	183	0.72	0.54	0.98	0.87	-62.30	0.72	0.54
RF00458_A	202	**0.78**	**0.64**	0.58	0.39	-49.90	**0.63**	**0.45**
RF00193_A	273	**0.86**	**0.73**	0.79	0.60	-73.60	0.55	0.38
RF00231_A	275	**0.97**	**0.71**	0.71	0.41	-89.00	**0.92**	**0.63**
RF00503_A	293	**0.95**	**0.87**	0.70	0.44	-55.40	0.84	0.63
RF00030_A	297	**0.70**	**0.53**	0.68	0.48	-83.77	0.46	0.29
RF00216_A	302	**0.63**	**0.46**	0.40	0.21	-117.24	0.49	0.29
RF00010_A	312	0.68	0.57	0.77	0.63	-117.70	0.67	0.54
RF00009_A	320	**0.77**	**0.35**	0.57	0.22	-87.40	0.34	0.13
RF00100_A	330	**0.75**	**0.58**	0.40	0.23	-102.90	**0.71**	**0.50**
RF00036_A	337	0.63	0.50	0.94	0.86	-116.04	0.63	0.49
RF00209_A	379	**0.77**	**0.54**	0.75	0.46	-139.10	0.63	0.38
RF00024_A	451	**0.86**	**0.53**	0.80	0.48	-215.20	0.73	0.41
RF00210_A	462	**0.91**	**0.69**	0.80	0.56	-175.50	0.74	0.51
RF00177_A	482	0.82	0.63	0.93	0.74	-239.50	0.72	0.50
Average		0.75	0.58	0.76	0.53		0.59	0.41

and energy computation algorithms. Also the approach can be used for identifying common motifs, i.e., pseudoknots, in a set of sequences - this can be used in case a segment prediction is given but it is not known the family or the genes to which it belongs. If significant structures are present in the segment as well as other members of the same family, this may indicate a possible relation to the family.

Although this is still work in progress, the results are very encouraging, especially considering that our method proved to be capable of improving the prediction of the whole sequence. A method such as the one proposed here has several advantages over conventional approaches on motif finding: (1) It does not require any prior knowledge about the sequences being analyzed, which makes it a very practical tool. (2) Because it lends itself to run in parallel, computational time will not be a critical issue. (3) It is very flexible and general in the sense that it can me combined with the prediction code of preference.

5 Future Work and Conclusions

This paper presents a method to rebuild RNA secondary structures from common motifs found in systematically sampled chunks of nucleotides. For 24 sequences of the 39 RNA segments with different lengths that we used for the validation, our method achieved results that are more accurate (either in terms of sensitivity or selectivity or both) than those achieved by the code we are using to predict the secondary structure of the sequence chunks when predicting the whole segments. The regression analysis outlined that (1) there is a significant relationship between the accuracy of the rebuilt structure (in terms of sensitivity and selectivity) and the sampling factors (i.e., window size, window step, and threshold values); (2) our method equally targets both the measurements of accuracy, i.e., sensitivity and selectivity; and (3) the minimum free energy cannot be trusted by itself as a quality measure of the secondary structure.

Current work includes the analysis of larger sets of longer RNA sequences and the use of other prediction codes for the chunks. As shown in [6], predictions can significantly benefit from the combined prediction capability of different codes as oppose to using single codes separately. We are also working on developing an intelligent strategy for generating chunks. As the results showed, the method is very sensitive to where the chunk begins and ends. One of our strategies to overcome this is to sample non-consecutive chunks. By doing so we can predict sequences that span over a longer number of nucleotides. We are also investigating the use of statistical methods for selecting the segmentation points. The idea is to train machine learning approaches on a set of sequences where the optimal cutting points have been identified a priori and use as attributes for this task the l-mers and inversions in the sequence.

Acknowledgments

This research is supported in part by NIH Grants No. 2S06-GM00812-37, 2T36-GM008789-04A1, and 5G12RR008124-11, and Texas Higher Education Coordinating Board grant 003661-0008-2006.

References

[1] M. Anwar, T. Nguyen, and M. Turcotte. Identification of consensus RNA secondary structures using suffix arrays. *BMC Bioinformatics*, 7:244, 2006.

[2] D. Ashlock and J. Schonfeld. Depth annotation of RNA folds for secondary structure motif search. In *Proc. of the 2005 IEEE Symposium on Comp. Intelligence in Bioinf. and Comp. Biology*, 2005.

[3] O. Bergig, D. Barash, and K. Kedem. RNA motif search using the structure to string str^2 method. In *Proc. of the 2004 IEEE Comp. Systems Bioinf. Conference*, 2004.

[4] D. Chew, K. Choi, H. Heidner, and M. Leung. Palindromes in SARS and other coronaviruses. *INFORMS J. Comp.*, 16:331–340, 2004.

[5] C. B. Do, D. Woods, and S. Batzoglou. CONTRAfold: RNA secondary structure prediction without physics-based models. *Bioinformatics*, 22(14):e90–e98, 2006.

[6] T. Estrada, A. Licon, and M. Taufer. CompPknots: a framework for parallel prediction and comparison of RNA secondary structures with pseudoknots. In *Proc. of 1st Frontier on High Performance Comp. and Networking Workshop*, 2006.

[7] B. Matthews. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta.*, (405):442–451, 1975.

[8] T. Nguyen and M. Turcotte. Exploring the space of rna secondary structure motifs using suffix arrays. In *Proc. of 6th Int. Symposium on Comp. Biology and Genome Informatics*, 2005.

[9] N. Pierce. NuPack: A software suite for the analysis and design of nucleic acids. http://www.nupack.org.

[10] J. Reeder and R. Giegerich. Design, implementation and evaluation of a practical pseudoknot folding algorithm based on thermodynamics. *BMC Bioinformatics*, 5:104, 2004.

[11] E. Rivas and S. Eddy. A dynamic programming algorithm for RNA structure prediction including pseudoknots. *J. Mol. Biol.*, 285:2053–2068, 1999.

[12] M. Taufer, M. Leung, K. Johnson, and A. Licon. RNAVLab: A unified environment for computational RNA structure snalysis based on grid computing technology. In *Proc. of the 6th IEEE Int. Workshop on High Performance Comp. Biology*, 2007.

[13] F. van Batenburg and et al. PseudoBase: a database with RNA pseudoknots. *Nucleic Acids Res.*, 28(1):201–204, 2000.

[14] M. Zuker. Computer prediction of RNA structure. *Methods Enzymol*, 180:262–288, 1989.

[15] M. Zuker, D. Mathews, and D. Turner. Algorithms and thermodynamics for RNA secondary structure prediction: A practical guide. In *In RNA Biochemistry and Biotechnology*. Kluwer Academic Publishers, 1999.